Prediction and Analysis of Functionally Important Sites in Protein Structure

A thesis submitted for the degree of

Doctor of Philosophy

by

Jaspreet Singh Sodhi

Department of Computer Science,

University College London

June 2005

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

Abstract

Identifying the molecular role of a target protein is becoming a pressing issue in structural bioinformatics. As the number of uncharacterised proteins accumulate and structural genomics initiatives pick up pace reliable and accurate approaches are required to make sense of the growing structural data in the most effective manner.

This study aims to improve the characterisation and cataloguing of functionally important regions in proteins to improve the transfer of information from structure to functional annotation. This is of great importance as effective and accurate annotations are a pre-requisite to unlocking the biological information encoded within the genomes of organisms.

A novel and automatic method is presented which identifies residues interacting with metal ions. The aim has been to focus the approach such that it is capable of the prediction of interactions even when reliable side-chain information is unavailable. The results demonstrate that a combination of sequence and structural features can be combined to achieve adequate predictions in both crystal structures as well as low resolution fold recognition models.

The findings of the metal binding study are used to extend the approach and predict larger interaction regions found in DNA binding proteins. These important

class of proteins are of particular interest given their fundamental control over biological processes. The results highlight effective classification is possible of both residues forming DNA contacts as well as the discrimination of DNA binding from non-DNA binding proteins. Crucially, DNA binding predictions are shown for a genome-wide study of *Sacchromyces cerevisiae*.

The final study of the thesis focuses the attention on the alternative problem of improving the quality of protein structure predictions from fold recognition. The results indicate that site classification provides an effective basis by which protein models can be assessed: this in turn leads to the development and benchmarking of a new method which significantly improves the quality of protein models.

Acknowledgements

I would like to thank my family and friends for providing the support and encouragement only they could provide.

I would also like to thank everyone I have worked with in the UCL Bioinformatics group. In particular: David Jones and Lorenz Wernisch for their supervision, feedback, and helpful discussions. Kevin Bryson, for productive discussion and feedback. Liam McGuffin, for encouragement, advise and collaboration which led to the work in the final chapter. Jonathan Ward, for helpful discussion throughout the course of the thesis. Tim Ebbles, for proof reading the fourth chapter and helpful comments. I would also like to acknowledge Liam McGuffin and Steffano Street for access to the distributed system.

This work was sponsored by the UK Medical Research Council Bioinformatics scholarship.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Understanding Protein Function

Identifying the functional properties of proteins, encoded within the genomes of organisms is a vital step in understanding how complex biological processes occur. Genome sequencing can only be regarded as the first step, identifying the function of proteins and their interactions presents a new set of challenges that will require the development of novel tools.

The definition of protein function, however, is often vague and can be described at varying levels from the molecular action through to physiological role. The molecular or biochemical role generally refers to a catalytic activity or binding property whilst, at a higher level, cellular function may be described by the metabolic path the protein participates in. The phenotypic function on the other hand relates to the contribution the protein exerts on the entire organism. Therefore, the interactions a protein participates in may be regarded as defining the molecular role that in turn dictates higher levels of function. Consequently understanding, characterising, and predicting protein interactions is an extremely valuable and important aspect in the post-genomic era.

An important point that should be considered is that the molecular role of a protein does not necessarily reveal the phenotypic function. A classic example is provided by the chymotrypsin family of serine proteases; in humans these proteins are found to be involved in a wide range of phenotypic functions ranging from food

digestion to a regulation function in the immune system, although the molecular function is closely related in both processes (Moult and Melamud, 2000).

Correctly determining the molecular role of a given protein can nonetheless provide valuable information that can be used to direct further research. A number of techniques and methods exist that can provide details regarding the molecular function of a protein. The literature reflects the growing interest in protein function analysis and prediction. A particularly important area is the analysis of protein structure to identify and characterise functional sites.

The purpose of this chapter is to highlight the important strategies for the analysis of protein structure to gain details of molecular function. To set this objective into context, an overview of various function classification schemes is initially presented. This is followed by a review of notable methods that aim to provide functional details given a protein structure. The final part of the chapter provides a synopsis of studies that have be performed in this thesis.

1.2 Function Classification

The wealth of genomic data that has become available in recent years has spurred the development of functional classification schemes. This is important as consistent terminology enhances the quality of database annotations as well as allowing functional comparisons to be accurately performed. Examples of different approaches and resources used to study protein function are discussed in the following sections.

1.2.1 Enzyme Commission

The Enzyme Commission (EC) (Enzyme Nomenclature, 1992) was one of the first examples of a structured approach for classifying protein functions. The motivation of the scheme was to provide a manageable framework within which the growing number of enzyme reactions being discovered could be categorised. The enzyme reactions are represented by a four number code where the first number provides information regarding the reaction class (such as oxidoreductase or ligase). The second number is dependent on the first and represents, for example, the type of substrate bound or the formation of a given bond type. The EC defines six top level classes of enzyme reactions, the ENZYME database (Bairoch, 1994) reveals that the six top level classes represent 63 and 215 distinct classes for the second and third EC levels respectively. The obvious limitation of EC classification is that it is designed

for enzyme reactions and is therefore inapplicable to other functional categories.

1.2.2 Gene Ontology

In recent years the Gene Ontology (GO) consortium (Harris et al., 2004) has aimed to provide a structured and controlled set of terms and classifications to improve the definition of protein function. The GO vocabularies (ontologies) were originally derived using information from FlyBase (Drosophila), the Saccharomyces Genome Database (SGD) and the Mouse Genome Informatics (MGI) resource. Recent additions have also included information for the Rat Genome Database. This should ensure adequate coverage of functional terms across a broad range of biological processes.

GO vocabularies describe gene products in terms of biological processes, cellular components and molecular functions and are organised as directed acyclic graphs (DAG). The important aspect of this structure is that child nodes in the graph can have more than a single parent as opposed to the single parent structure of hierarchies.

A number of different databases now employ GO terms to improve annotation quality and querying. For example, Enzyme Commission (EC) numbers can be matched to various levels of GO terms as well as can UniProt and Pfam entries. The ongoing uptake of GO terms is therefore likely to significantly impact the area

of protein function characterisation.

1.2.3 InterPro

The growing number of different databases and classification schemes can be difficult to effectively interpret. The InterPro resource (Mulder et al., 2002) was developed in order to provide a consensus of protein family information from a number of different sequence based databases. The system incorporates protein family and functional site information from several secondary databases including PROSITE (Falquet et al., 2002), PRINTS (Attwood et al., 1997, 2000), BLOCKS (Henikoff et al., 2000) and Pfam (Bateman et al., 2000). Although the combination of these methods within InterPro provides a powerful analysis tool, significant interpretation of the results is often required and as a result is not generally appropriate for large scale automatic analysis.

The disadvantages of sequence motif based methods in general is that they rely on a given motif or signature having been observed before and assigned to a function. This often requires many examples, supplemented with experiment, and significant human intervention. However, the inherent limitation with all sequence based approaches is the fact that they are based in one-dimension. As such sequence based methods are unable to directly encode 3D spatial organisation of functional residues and the atoms responsible for biochemical action in the folded protein. In many

cases the greater level of detail provided by protein structure is therefore likely to be more appropriate for functional characterisation.

1.2.4 Protein Sequence Databases

The SWISS-PROT database (Bairoch and Apweiler, 1997) provides detailed functional information that has been manually curated for known protein sequences. The information is derived either by experimental methods or through homology based approaches and includes descriptions of protein function as well as links to other resources. The TrEMBL database supplements SWISS-PROT providing an automated system where translations of nucleotide sequences from EMBL are annotated.

More recently the UniProt system was presented (Apweiler et al., 2004) to provide an accurate and rigorous system for high quality annotations. Manually curated information from SWISS-PROT is included as well as automated annotations within TrEMBL. UniProt also provides several non-redundant protein sequence databases which can be queried to decipher relationships to protein sequences of interest. Resources such as UniProt are extremely important to allow the research community to easily obtain accurate, well annotated protein information.

1.2.5 Homologous Clustering

Identifying common features between sequences provides a powerful indicator of evolutionary and functional relationships. The Clusters of Orthologous Groups (COGs) (Tatusov et al., 1997) approach compares the genomes of different organisms and groups together orthologs; equivalent genes in different species with common ancestry. The fact that orthologs typically share the same function allows one to transfer functional information to new members of a given COG. However, correctly identifying orthologous relationships in the first place is not always a trivial task and as such may not always be reliable.

1.3 Analysis of Protein Structure

Understanding the molecular role of a protein can be trivial if a close homolog has been characterised. However, in the absence of reliable sequence similarity, protein structure often provides more distant evolutionary details. For example, identifying the overall organisation of the polypeptide chain can allow fold similarities to be identified as well as aid the discrimination of buried core residues from surface residues. Alternatively, bound ligands may highlight catalytic residues and active sites, thereby providing clues of the mechanism of function. Therefore methods and resources which provide a means to analyse and compare protein structure offer

a very powerful alternative to sequence comparisons. Several aspects of protein structure analysis are discussed in the following sections.

Before considering approaches for the analysis of protein structure it is appropriate to consider the methods that provide structural details of proteins. A discussion of the computational approaches for protein structure predictions is covered in Chapter 5, for the current discussion the focus will be on experimental methods.

1.3.1 Experimental Structure Determination

Currently there are two main experimental techniques for accurately discovering the 3D-structure of a protein: X-ray crystallography and nuclear magnetic resonance (NMR) spectroscopy. The initial step for both methods requires significant quantities of the target protein. This is achieved using molecular biology techniques to amplify a target gene, clone, express and finally purify the protein. This process may be regarded as the limiting step in protein structure determination as it can often become a very laborious and time consuming exercise. Further, expression of eukaryotic proteins is generally more difficult as compared to prokaryotic organisms, requiring perhaps specialised cellular machinery, post-translational modifications or chaperone proteins. A brief overview of X-ray crystallography and NMR is presented below.

X-ray Crystallography

The process of crystallising a macromolecule, although significantly improved in recent years, is still a time consuming and challenging task. The physical properties of a protein are not generally ideal to form the regular lattice of a crystal. Prohibitive factors include the irregular shape of a folded protein, the requirement of active site substrates, conformational changes, weak interactions between asymmetric units of the protein and the degree of solvation. Consequently many different strategies are used in order to grow and maintain fragile protein crystals, including varying pH conditions, addition of precipitant salts as well as inclusion of substrates or inhibitors. As a result high throughput approaches often implement the use of robots to optimise and screen crystallisation conditions. An important breakthrough was reported by Teng (1990) which involves cryo-freezing protein crystals to reduce damage caused by the high intensity X-ray beam.

The advantages of the crystallographic approach for structural determination are mainly the high degree of accuracy and the ability to process large macromolecules. However, crystal structures represent a time-averaged representation of the protein structure and therefore does not provide details regarding the dynamic nature of the molecule.

NMR Spectroscopy

NMR is generally less time consuming as compared to X-ray crystallography and can be performed directly on samples of the purified protein without the need for crystallisation. The method essentially involves measuring the magnetic spin of atomic nuclei to determine the distance between atoms. Analysis of the resultant data provides the 3D-atomic structure of the protein. The major limitation with NMR is that high resolution atomic information is not possible for proteins comprising more than a few hundred residues or >23-30 kDa (Branden and Tooze, 1998).

1.3.2 The Protein Data Bank

The Protein Data Bank (PDB) (Berman et al., 2000) provides a repository for protein structures obtained mainly from crystallography or NMR. The rapid growth of the PDB, since it was first introduced in 1972, has been well documented and ongoing updates are available at the Research Collaboratory for Structural Bioinformatics (RCSB). From the years 2000 to 2003 a staggering 14,204 new protein structures were deposited to the PDB. The year on year increases have resulted from both improvements in experimental techniques as well as structural genomics initiatives gathering pace.

The growth of the PDB is also reflected in the number of protein folds. Analysis of the hierarchical classification system SCOP (Lo Conte et al., 2000) shows that, in

August 2003, 2327 protein families corresponded to 800 unique protein folds. However, during the 2000-2003 period only 236 new folds were deposited as compared to the 14,204 new structures deposited. This highlights the conserved nature of protein structure, indicating a limted number of folds. Estimates for the number of protein folds have varied significantly over the years, Chothia (1992) proposed that there may only be 1000 protein folds. However, it may be the case that if gaps in protein structure space are to be filled, innovative experimental breakthroughs will be required.

1.3.3 Hierarchical Classification

Hierarchical structural classification methods aim to group proteins according to varying levels of structural similarites. The premise of such approaches is that protein structure is more conserved as compared to sequence and can therefore allow subtle evolutionary relationships to be identified (Orengo et al., 1999). In general the top of the hierarchy describes the overall secondary structure of residues in the protein. Protein structures can be described as mainly α, mainly β or $\alpha - \beta$. Lower levels of the hierarchy generally describe the organization of secondary structures and their arrangement in the folded protein.

The SCOP (Andreeva et al., 2004), CATH (Orengo et al., 2002) and DALI (Holm and Sander, 1994) databases are popular examples of resources that contain group-

ings of protein structures. These systems use varying strategies: the SCOP method is a manual based approach whereas DALI is fully automated. The CATH system uses both automatic and manual strategies in the grouping procedure. Analyses between SCOP, CATH and DALI has shown agreement >80% at the super-family level with most differences due to the automation (Hadley and Jones, 1999).

1.3.4 Inferring Function from Fold Relationships

Details of a protein's fold can provide important clues to its likely function. This is especially useful when little else is known about a given protein sequence. However, an important distinction that needs to be made early on is whether a protein structure is homologous (common ancestry) to a characterised structure or analogous (no common ancestry). Analogous proteins may arise simply due to physiochemical constraints on protein folding (Orengo et al., 1999) and may not be functionally comparable.

Consequently, knowing a protein's fold does not necessarily reveal its function. This has been strikingly illustrated by the occurrence of 'superfolds' (Orengo et al., 1994), commonly occurring protein fold configurations that can span significantly different functional classes. The TIM-barrel proteins provide an example of these functionally promiscuous folds having been shown to span 60 different enzymatic functions (Nagano et al., 1999). Interestingly many 'superfold' proteins have been

shown to contain preferred binding site locations known as 'supersites' capable of binding varying substrate types (Russell et al., 1998).

The alternative scenario can occur through the process of convergent evolution, whereby proteins have evolved to perform similar biochemical functions independently. The classical example of this process is exemplified by chymotrypsin and subtilisin serine proteases, both these families perform similar functions using similar catalytic strategies although the overall folds of the proteins are completely different.

1.4 Structural Genomics

Sequence comparison algorithms allow a target gene or protein sequence to be analysed against sequence repositories. However, novel gene products are unlikely to share sequence features which can be reliably detected. For these growing numbers of cases the information provided by protein structure may be the only means to understand the molecular role of the target. It may be that, in a folded state, key active site regions become discernible allowing matches to other protein structures sharing the same functional site region and therefore function. The sequence-structure-function paradigm is therefore an important route by which newly discovered protein sequences can be analysed and characterised.

The aim of structural genomics initiatives is to determine the structure of representatives of all protein families within target genomes. This is required to bridge the gulf between the number of known protein sequences and the number of solved structures. Obtaining detailed structural information is likely to uncover important features for proteins of unknown function.

1.4.1 Challenges

The Protein Structure Initiative (PSI) has identified ambitious goals in order to bridge the gap between the sequence and structure knowledge base. The motivation behind this is that protein structure is more conserved, as compared to sequence space, and is therefore more likely to provide greater functional and evolutionary details of novel protein sequences. However, the fact that identical protein folds can perform varying functions (analogous proteins) has resulted in a fundamental paradigm shift in structural biology: structure determination is no longer a guarantee of elucidating protein function. Consequently, there are growing numbers of structures available for functionally uncharacterised or hypothetical proteins.

1.4.2 Progress and Future Direction

Much of the planning and development of the PSI took during the late 1990's, however it was during 2004 that the second stage of the initiative was approved

(Nature Editorial, 2004). This encouraging announcement indicates the importance of the proposed task as well as the potential benefits. It is hoped that an estimated 10,000 structures will be solved over a ten year period from the various structural genomics centres.

The genome of *Mycoplasma genitalium* was the basis of one of the earliest structural genomics project initiated to bridge the gap between protein sequence and structure data banks. Currently, the PDB provides an updated resource which provides information for many different structural genomics projects. Table 1.1 provides an overview of a selection of these projects (targeting the genomes of *Mycoplasma genitalium, Mycoplasma pneumoniae, Arabidopsis thaliana, Thermotoga maritima, Caenorhabditis elegans) and their status in the latter half of 2004. Analysis of the PDB shows that during 2003, 176 structural genomics targets were submitted to the PDB. Before the end of 2004, this number had risen to 534 new structures (Westbrook et al., 2002). Although this is an ever changing situation, this snap shot of the progress which has already been made illustrates the urgent need for effective tools and techniques to analyse structural data.*

SG Centre Name	*Target Genome*	*Solved Structures (in PDB)*	*Targets Selected*
Berkeley SGC	*M. genitalium, M. Pneumoniae*	48	911
Centre for Eukaryotic SG	*A. thaliana*	32	5630
Mid-west Centre for SG	*T. maritime, C. elegans*	133	10,952

Table 1.1: Overview of a selection of Structural Genomics Initiatives in 2004.

1.5 From Structure to Function

As discussed, structural insights often provide a means of identifying distant evolutionary relationships as well as providing functional information which may be undetectable at the sequence level. This is particularly important as structural genomics initiatives are rapidly solving protein structures for novel protein sequences resulting in a pressing need for reliable and efficient methods for bridging the gap between structure and function.

Therefore the analysis of protein structure, ranging from atomic level, through to overall domain organisation, is a powerful approach for identifying functional characteristics which can be applied in a predictive manner. In many cases, key functional regions in protein structure may distinguish functional activity allowing, for example, enzyme active sites or DNA binding regions, to be uncovered.

The rapid growth of the Brookhaven Protein Data Bank (PDB) (Berman et al., 2000) however highlights several challenges. Gene products from different species may exhibit similar biological function, but show little or no sequence similarity due to convergent evolution. Structural classification of protein domains such as CATH (Orengo et al., 2002), SCOP (Andreeva et al., 2004) and FSSP (Holm and Sander, 1994) reveal that members of the same structural family can span different functional classes. Furthermore, key active sites may be conserved although there is little overall structural and sequence similarity. It is therefore clear that analysis of functional regions will not only allow the development of more reliable genome annotation but also enhance our knowledge of the biological role of proteins at a cellular level.

Figure 1.1 summarises key strategies for gaining functional information given a protein structure: important areas include identifying interaction sites (Casari et al., 1995), comparing the overall fold topology (Andreeva et al., 2004), observing and analysing binding pockets (Fetrow and Skolnick, 1998) or cavities (Laskowski, 1995) as well as identifying bound ligands (Wei and Altman, 2003).

Given the detailed information that may be gleaned from the analysis of protein structure, numerous studies and tools have been reported in the literature which attempt to characterise functional properties of proteins of interest.

A detailed discussion of methods for the study of protein-metal and protein-DNA

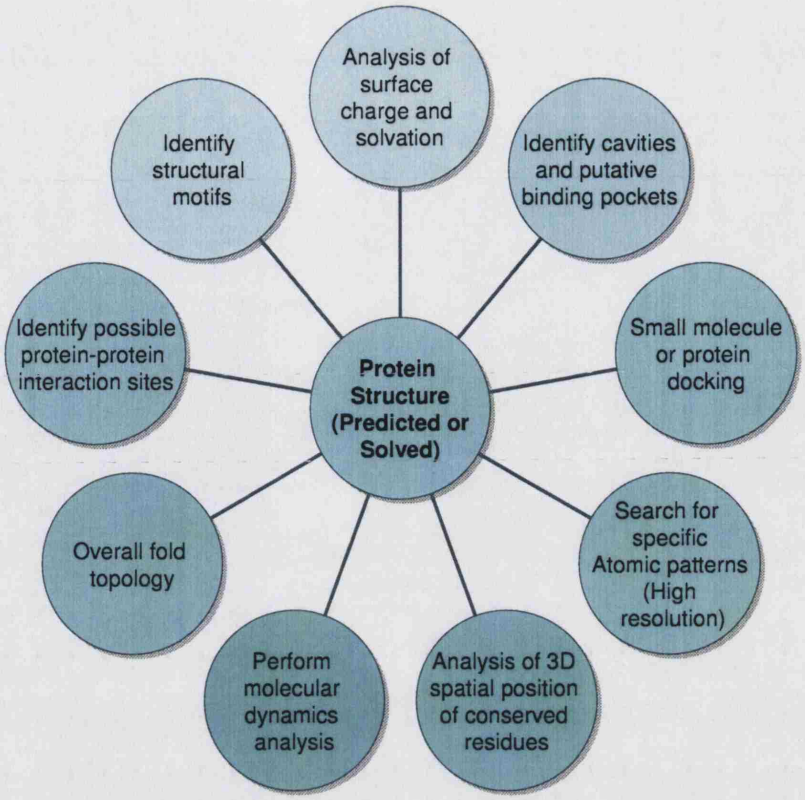


Figure 1.1: Strategies for gaining functional information from protein structure

interactions is included in chapters three and four where these types of sites have been investigated in greater detail. The goal of this section is to present a selection of different methods and strategies which utilise observations and trends in structure and sequence to bridge the structure-function gap.

1.5.1 Classifying Functional Sites

Identifying spatial motifs, interaction interfaces and binding sites can significantly aid the elucidation of a protein's molecular role. Several groups have developed

atomic level methods for identifying and characterising functional site regions (Artymiuk et al., 1994; Wallace et al., 1997; Zhang et al., 1998; Kleywegt, 1999). Although the implementation of these approaches vary, they all share a common theme. The methods incorporate information, in one form or another, obtained from structural analysis of proteins.

TESS

The Template Search and Superposition (TESS) method of Wallace et al. (1997) provides an atomic level based description of active site regions. Atomic coordinates of a given PDB file are transformed to a common reference frame, and the grid positions of all atoms within a cut-off are recorded in a hash table. This is repeated for all the structures in the PDB. A geometric hashing algorithm is then applied to rapidly match a query template to the stored information from which the root mean square deviation (RMSD) is calculated.

The TESS approach was originally used to develop a template for the serine proteinases and ribonucleases. The Procat database (Wallace et al., 1997) is an online server that offers access to such templates, although at present it is limited in the number of templates available. The atomic templates used in TESS are dependent on the accurate placement of side-chain atoms, although some variation is allowed by incorporating a distance cut-off for surrounding grid point. Consequently

the TESS method will most likely be unsuitable for poorly resolved or modelled structures where side chain atoms are likely to have relatively high RMSDs. TESS also lacks sequence information, and therefore any useful evolutionary data may be missed as well as higher order structural data which is also omitted.

FEATURE

An interesting algorithm, FEATURE, has been described by Wei and Altman (2003). Essentially the method aims to build a statistical model of a specific type of functional region given a set of site and non-site examples. The method has been presented for calcium site detection (Wei and Altman, 1998) and was more recently presented for larger ATP binding sites (Wei and Altman, 2003).

The algorithm works by dividing the region around the site of interest into concentric shells of 1Å thickness then recording several details within each of these sub-volumes. The properties recorded range from chemical information of atoms through to amino acid type, accessibility as well as structural details. Numerical values represent each property within the sub-volume and form the basis of the distribution describing the site. Distributions from sites and non-sites are compared and the statistical significance is assessed. Interestingly, the method does not directly encode directional information but only records the shell position of the individual features.

FEATURE was subsequently improved to calculate a log-odds score based on Bayesian analysis and was applied to the testing of query structures to determine calcium binding ability (Liang et al., 2003b). The method was also used to assess the prediction of sites in modelled structures (Wei et al., 1999). However, the conclusions of this study showed that even small deviations in atomic positions, in the region around the calcium ion, would lead to misclassification even though the overall RMSD of the protein models were low. It should be noted that the FEATURE method is based solely on atomic information and lacks sequence details, therefore the approach does not contain any direct evolutionary knowledge. Nonetheless, the approach offers an interesting perspective to functional site analysis.

The Web-feature tool (Liang et al., 2003a) was presented more recently, providing online access to the FEATURE algorithm. The resource requires as input a 3D structure together with a set of sites of interest, from which a statistical model is generated. A Bayesian based scoring system is used to identify important structural and atomic properties from the training data of sites and non-sites provided by the user. Therefore the system does require significant manual processing and relies upon adequate examples for the region of interest.

SPASM and RIGOR

The SPASM (Spatial Arrangement of Side and Main chains) (Kleywegt, 1999) algorithm implements a recursive depth-first search to match a user defined spatial motif, represented in PDB format as a set of residues of interest, to a database of structures. The aim is to match residue types, and distances between the residues, of the input motif with a set of protein structures of interest. Such a methodology is useful in that users have the control to generate a spatial motif they deem to be of significance, however, this may require in-depth knowledge which could in turn limit the applicability of the method.

The approach also offers the advantage of being insensitive to errors and ambiguities in atomic positions as only the positions of Cα (or pseudo centre for Gly) atoms recorded. The drawback of this 'fuzziness' is an increased tendency for false negatives.

The RIGOR program implements the reverse process to that of SPASM, allowing query proteins to be scanned against a database of pre-compiled motifs. RIGOR limits the residues which are scanned to only charged and polar residues as well as residues in contact with heterogeneous groups.

A drawback of the system is the measure of similarity between a motif and query region of interest. The correctness of a match is determined by the root mean square deviation, however, the statistical significance a match is not evaluated. Another

drawback of the methodology of SPASM and RIGOR is that no form of sequence information is included in the approach. Although this may be appropriate for truly novel sequences it is quite often the case that some sequence information, however distant, is obtainable.

Fuzzy Functional Forms

The 'Fuzzy Functional Forms' (FFF) method, originally presented by Fetrow and Skolnick (1998), is a powerful technique which combines sequence and structural information to allow the identification of functional sites. In addition the algorithm utilises information from the literature to define a set of constraints capable of uniquely describing the region of interest.

This approach was first developed for the disulphide oxidoreductase activity of glutaredoxin and thioredoxin (Fetrow and Skolnick, 1998) and later extended to a genome-wide anaysis of *E. coli* (Zhang et al., 1998). Interestingly the method was successfully applied to modelled structures from *ab initio* and threading methods by loosening the constraints of the functional template.

Although the approach provides a seemingly powerful tool for functional characterisation at the genome scale it requires information to be collated from several sources. The site to be analysed must have been previously described and solved to initially develop the FFF. Also, the method has not been tested for classification of

non-enzyme functions.

The FFF was more recently updated (Cammer et al., 2003) to allow application to a larger set of site types in a more efficient manner. The system was updated to incorporate 193 enzyme active sites for all 6 EC classes. Multiple sequence alignments were used to identify conserved positions followed by grouping neighbouring residues in space. The approach was shown to be effective for the subclassification of protein families using the protein kinase family as an example.

1.5.2 Evolutionary Trace

Analysing the conservation of residues within protein families is a powerful approach to aid the detection of functionally important regions. Quite often conserved regions will be associated with a sequence or structural motif implicating a particular function. Multiple sequence alignments highlight completely conserved position within a protein family. However, subtle patterns of conservation between different protein families can provide important information regarding functional regions. Residues which are completely conserved within a subfamily but differ between subfamilies are known as tree-determinant positions and often allow subtle functional relationships to be uncovered (Casari et al., 1995).

The evolutionary trace approach (Lichtarge et al., 2003; Yao et al., 2003) uses the trends between conserved and variant positions within and between protein sub-

families to predict active sites and functionally important regions. Originally this approach required manual clusters of protein families to be constructed based on sequence similarities and functional characteristics. Tree determinant residues are then mapped onto the protein structure. An important development in the evolutionary trace approach was automation of the technique as well as providing a quantitative basis to assess the significance of overlap with actual functional sites (Yao et al., 2003). In addition a comparison of methods to identify the statistical basis of tree-determinant residues was presented by del Sol Mesa et al. (2003). The conclusion of this study highlights that tree determinant positions are likely to become apparent as more members of the subfamily are used in the multiple alignment.

1.5.3 Locating Surface Clefts

A particularly useful approach to identifying possible functional sites when given a protein structure is to locate surface clefts. Measuring the volumes and shape characteristics of such clefts is often a powerful indicator for functionally relevant parts of the protein. This was demonstrated by the SURFNET program, developed by Laskowski (1995), which illustrated that, for enzymes, the largest clefts generally correspond to the ligand binding site. A powerful extension of cleft detection is to combine residue conservation results, as clefts which comprise conserved residues

are more likely to be functionally important (Laskowski et al., 2003).

1.5.4 Graph Theoretic Approaches

Graph theoretic approaches have been shown to provide an effective approach for matching and identifying functionally important regions in protein structure (Artymiuk et al., 1994). The premise behind such techniques is to represent protein structures, or parts of structures, as graphs. For example, atoms and inter-atomic distances may form graph vertices and edges respectively. However, due to the size of protein structures, identifying matching regions or sub-graphs, known as clique detection, between two proteins is computationally challenging.

Nonetheless the approach offers many advantages; firstly proteins may be queried rapidly against a database of functionally important regions allowing for automation. However, the limitation is that functional regions need to have been documented previously and high resolution structural data is often required.

Schmitt et al. (2002) recently presented a graph based approach to compare residues flanking clefts in protein structures. Residues were represented using a simplified system whereby the physiochemical properties could be matched rapidly using a clique detection algorithm. A database, containing such representations, for known binding sites was queried given a new site region. The approach was shown to retrieve protein cavities accommodating similar or related ligands or proteins with

related catalytic mechanisms.

A similar approach was presented by Kinoshita and Nakamura (2003) who compared protein surfaces against a database of molecular surfaces, the eF-site (electrostatic surface of Functional site) database. Electrostatic potentials of surface active sites, phosphate sites and antigen binding sites are stored alongside PROSITE motifs. A clique detection algorithm allows site matches to be identified and was shown to be effective at correctly identifying sites in different protein folds.

Gardiner et al. (1997) compared various commonly used algorithms to locate the maximum common sub-graph (MCS) between graph representations of molecular structures. The graph representation used in this study defined α helices and β sheets as the graph vertices whilst distances between the secondary structure elements formed the graph edges.

The study demonstrated that the clique detection algorithm by Carraghan and Pardalos (1990) is generally two to three times faster than the commonly used Bron-Kerbosch algorithm (Bron and Kerbosch, 1973). However, the advantage with the Bron-Kerbosch algorithm is that it will return all cliques above a given threshold whereas Carraghan and Pardalos approach only identifies the maximum clique. The authors therefore suggest a combination of the two methods as a valuable strategy depending on the requirements of a search.

1.5.5 PINTS

A common means to quantify the structural similarity between parts of proteins has been to calculate the root mean square deviation between sets of atoms from each protein. Unfortunately RMSD comparisons are not comparable if the number and type of atoms are different. Stark et al. (2003) presented the PINTS (Patterns In Non-homologous Tertiary Structures) method which aimed to provide a statistical significance of the RMSD between regions of protein structures. The approach uses a geometrical model which allows similarities in the spatial arrangements of amino acids to be assessed. The approach however does not aim to provide a means to measure fold similarities but geometrical similarities within sub-regions of the protein.

The Stark paper presented two varying models for assessing residue positions. The first, the independence model, assumes that single atoms from residues are randomly and independently distributed in space. Comparisons were performed between patterns consisting 2-8 Cα atoms and a background database. However, assuming independence between residues is not always appropriate, for example for functional site regions. The authors of the PINTS method therefore also presented a modified dependence model whereby the number of atoms included in the search pattern for each residue was successively increased.

The PINTS method is accessible via a web-server allowing protein structures to

be compared against a database of patterns. Conversely a structural pattern, up to 100 amino acids can be queried against a database of protein structures.

1.5.6 The ProFunc Integrated Resource

Recently the ProFunc system has been reported by Laskowski et al. (2003) which automatically performs several different analyses, given a protein structure, to provide clues as to its probable function. The system identifies conserved residues, using CLUSTALW, and locates clefts in the structure using SURFNET (Laskowski, 1995). Conserved residues can be viewed on the protein structure and superimposed with the SURFNET results to isolate conserved pockets, often highlighting a functional site. In the final stage of ProFunc, the Jess program (Barker and Thornton, 2003) is applied to screen the query structure against a set of 189 enzyme templates. Although the ProFunc system is continually being developed, the system already provides a very useful tool for the analysis of new protein structures.

1.6 Synopsis of Studies

The undertaking presented in this thesis has been to provide a means to combine varying information sources for characterising and predicating functional regions in protein structure. This has led to the development of a relational database system in

chapter two which combines evolutionary sequence features, in the form of sequence profiles, for regions of functional importance along with structural observations. Analysis of the database has provided useful information on features present in metal sites commonly found in biological macromolecules. Analysis of sequence features from PSI-BLAST profiles highlights the greater evolutionary constraints on metal binding residues whilst analysis of SCOP superfamilies highlight dominant fold types which utilise metals.

The development of a novel tool to automatically detect and classify metal binding site residues is subsequently presented in chapter three. The focus has been to develop classification techniques, utilising sequence and structural features, that can be derived from moderate quality protein models. Of particular importance has been the development of an effective confidence estimation approach to allow reliable discrimination between different classifiers. The findings provide clear evidence that effective detection of metal interacting residues is possible in both crystal structures as well as moderate quality fold recognition derived structural models.

In chapter four, the metal site classification scheme is extended to larger interaction sites found in the interface region between proteins and DNA. The fundamental importance of these classes of proteins is highlighted by their significant presence in the genomes of complex organisms. Processes such as gene regulation, DNA repair and replication are vital for maintenance of cellular processes. The initial analyses

aim to assess the ability to identify DNA binding site residues. Subsequent work has involved the discrimination of DNA binding proteins as compared to a non-DNA binding decoy set. Enhancing the quality of genome annotations forms the basis of the final part of the chapter. Site detection is assessed for structures predicted using a distributed version of the well established GenTHREADER fold recognition method. The DNA classification tool is validated on a set of DNA binding proteins of the *S. cerevisiae* genome annotated as DNA binding from the Gene Ontology (GO). Finally, an analysis of proteins described as having unknown functions reveals a number of significant hits for DNA interactions.

The final research chapter focuses on developing the findings of chapter three to incorporate site predictions into a new tool to enhance the scoring of protein models from fold recognition. The aim of the chapter has been to provide a better correlation between fold assignment and model quality as compared to the current system incorporated within GenTHREADER. Inclusion of site predictions along with a model quality checking parameter and a secondary structure alignment metric is benchmarked against the original inputs used in the GenTHREADER method. The results demonstrate a statistically significant improvement in model quality is attainable by incorporating site predictions along with the additional new and original inputs of GenTHREADER. However, the findings suggest that the majority of improvement results by training to optimise model quality directly. Independent

validation on LiveBench (Rychlewski et al., 2003) is presented, highlighting the improvement in identifying models of higher quality.

Chapter 2

FuncSite: A Database to Combine Sequence Profile and Structural Information for Functional Regions

2.1 Introduction

Recent years have seen a significant paradigm shift in the biological sciences. This has resulted due to the explosion of the amount of highly complex and varied data from sequencing projects, improvements in experimental techniques as well as the introduction of new technologies. Effectively managing the rapidly growing and diverse data in the biological sciences has therefore become an important challenge. The cataloguing, storage and analysis of such data will allow more pertinent questions to be formulated and provide a basis from which complex patterns can be understood.

An important aspect of the challenge is the requirement for well designed and easily accessible resources which will improve both the management and mining of the data. Dedicated publications providing an overview of web-based resources (Galperin, 2004) highlight the growing number of systems which aim to provide improved access to information, to drive forward the development of tools and hypotheses. A brief overview of a selection of resources providing sequence, structural or functional information and analysis is presented below.

2.1.1 Sequence Repositories

An important challenge has been providing an effective means to organise and maintain DNA sequences from the numerous genome sequencing projects worldwide. The GenBank resource (Benson et al., 2004) provides a comprehensive database which contains DNA sequences from over 140,000 organisms from both large sequencing projects as well as individual laboratories. Data is exchanged between GenBank, EMBL and the Data Bank of Japan to ensure information is updated in a timely manner worldwide. The GenBank resource is updated twice a month and can be accessed via the NCBI homepage (*http : //www.ncbi.nlm.nih.gov*).

More recently the UniProt system was presented (Apweiler et al., 2004) to provide an accurate and rigorous system for high quality annotations. Manually curated information from SWISS-PROT is included as well as more automated annotations within TrEMBL. UniProt also provides several non-redundant protein sequence databases which can be queried to decipher relationships to protein sequences of interest. Resources such as UniProt are extremely important to allow the research community to easily obtain accurate, well annotated protein information.

2.1.2 Sequence Analysis

Sequence searches are generally the initial analyses performed to evaluate ancestral relationships between sequences, indicating functional and structural similarities. A

sequence identity $\geq$30% generally indicates some degree of significance, however for the twilight zone ($\leq$25%-30%), the distinction between significant and insignificant similarity is often difficult (Orengo et al., 2003).

The statistical evaluation employed by the BLAST2 (Altschul et al., 1990) algorithm provides a robust measure to assess the significance of a sequence alignment. The method calculates an expectation value (E-value) of obtaining a given alignment score or greater by random chance. Importantly the E-value takes the size of the sequence database being screened into account.

2.1.3 Gene Ontology

The Gene Ontology (GO) aims to provide a structured and controlled set of terms and classifications which range from the molecular level through to biological process. The vocabularies are continually updated and reviewed by a large consortium of researchers. This is an important aspect as the quality of data in the biological sciences can vary tremendously and it is not always easy to decipher or interpret annotations. An important aspect of GO has been to provide focused annotations as well as more general descriptions which provide a intuitive feel for protein functions. GO vocabularies were originally derived using information from FlyBase (Drosophila), the Saccharomyces Genome Database (SGD) and the Mouse Genome Informatics (MGI). The GO database currently contains a wealth of information

from a large number of different projects as well as linking to various other information resources such as Enzyme Commission (EC), UniProt and Pfam. Resources such as GO are likely to prove tremendously useful if high quality, rigorous analyses of the growing data in biological sciences is to be possible.

2.1.4 The Structure Data Bank

The protein data bank (PDB) (Westbrook et al., 2002) contains a wealth of structural information derived mainly from X-ray crystallography as well as NMR. Improvements in experimental techniques, as well as structural genomics initiatives, has resulted in exponential growth of the PDB. Tools to effectively manage the information provided by protein structure are therefore of great importance.

2.1.5 Structure Analysis Tools

Structural Classification

Structural classification resources such as FSSP (Holm and Park, 2000), SCOP (Lo Conte et al., 2000) and CATH (Orengo et al., 2002) provide important structural details of proteins. The databases group protein structures at varying levels, for instance to a particular fold type or superfamily. This allows distant evolutionary relationships to be identified which are not detectable from sequence based approaches.

PDBsum

Several tools are available to analyse sequence and structural features of proteins. The PDBsum (Laskowski, 2001; Laskowski et al., 1997) web-based service provides detailed information for entries in the PDB. Interacting residues are highlighted, as are interactions with DNA/RNA, ligands and metal ions. Ligand molecule interactions are presented schematically using the LIGPLOT program (Wallace et al., 1995) whilst interactions to DNA molecules can be viewed using the NUCPLOT (Luscombe et al., 1997) program.

Another useful feature included in PDBsum is the ability to view PROSITE patterns mapped to the protein structure. The PDBsum tool is therefore a powerful resource providing an overview of features of a given PDB entry, as well as linking to many other resources.

Web-Feature

The Web-feature tool (Liang et al., 2003a) provides online access to the FEATURE algorithm which allows users to scan query structures for functional sites. The resource requires as input a 3D structure together with a set of sites of interest, from which a statistical model is generated. A Bayesian based scoring system is used to identify important structural and atomic properties from the training data of sites and non-sites provided by the user.

Catalytic Sites Atlas

Although the PDB does contain SITE records which relate to residues making up important catalytic sites the information is often incomplete and unreliable. The Catalytic Site Atlas, developed by Porter et al. (2004), includes hand-curated annotations for enzyme catalytic residues. Information is extracted from the literature as well as homology analysis. Sites can be searched using SWISS-PROT identifier as well as EC number or PDB code. The original version of the system contained 177 manually curated entries and over 2600 homologous entries.

SWISS-MODEL

A database of three-dimensional comparative models, generated by the fully automated SWISS-MODEL method, is available in the SWISS-MODEL repository (Kopp and Schwede, 2004). Over 300,000 models are contained in the database for sequences present in both SWISS-PROT and TrEMBL. Crucially, regular updates mean new template structures can be incorporated into the system thereby providing a reliable and updated source of model predictions. Several aspects of the models can be viewed including the quality of the prediction, alignment to template as well as links to SWISS-PROT.

SURFACE

The SURFACE (SUrface Residues and Functions Annotated, Compared and Evaluated) database (Ferre et al., 2004) provides a useful repository of protein surface patches indicating putative functional sites. Patches are annotated with structure and sequence information as well as information regarding function and possible interactions. The system also allows comparisons between proteins allowing quantitative similarities to be calculated. A graphical representation of surface patches as well as the annotations is available though a web interface.

PROMISE

The PROMISE (prosthetic centres and metal ions in protein active sites) database (Castagnetto et al., 2002) provides a useful resource which combines various types of information for metal containing proteins. An important feature of this system is the ability to view geometrical details of metal co-ordination. The system extracts first and second shell interaction information from crystal structures as well as recognising the type of site, the number and composition of metal ions present. The database provides details of high resolution information encapsulating atomic features of metal binding regions.

2.1.6 Chapter Overview

With the aim of developing our understanding of functional regions in proteins, we present the FuncSite database combining key structural and evolutionary information in a relational database model. Sequence conservation of functional regions are captured and stored from PSI-BLAST position specific score matrices (PSSM) whilst structural information is derived either directly or indirectly from co-ordinates contained in the publicly available structural database. The underlying aim we adhere to is not to be too dependent on specific placement of side chain atoms as we wish to develop a system capable of functional analysis using structural models, where high resolution features may not be available or reliable.

The design of FuncSite is discussed and several results are presented for commonly occurring metal ion sites. For the purposes of this chapter the discussion is focused on calcium containing proteins, however, the aim of the system is to provide a general framework within which sequence profile and structural information can be combined. The results provide consistent findings with known structural features of calcium sites as well as interesting features of residue conservation. The final part of the chapter provides details of commonly observed metal containing superfamilies.

2.2 Database Design and Development

The postgres SQL relational database management system was used to develop the FuncSite database. The August 2003 release of the PDB, containing 21,862 protein structures was used. The underlying aim in the development of FuncSite database schema was primarily to seamlessly merge site information derived from various sources. The design of the system is discussed in more detail in the following sections.

2.2.1 FuncSite Design and Development

The overall database schema is illustrated in Figure 2.1. The main information contained within FuncSite is derived from the PDB, DSSP and PSI-BLAST score profiles. Structural annotation is included from SCOP. The overall organisation of the FuncSite database model will be discussed in the following sections.

2.2.2 Hetero Groups Table

The Java program, *HetLoader*, was developed to parse and extract heterogeneous (Het) group information from the PDB files and load the data into the *HetGroup* table within FuncSite. This was achieved using Java Database Connectivity (JDBC). Each record provides information about the Het groups for a particular PDB file.

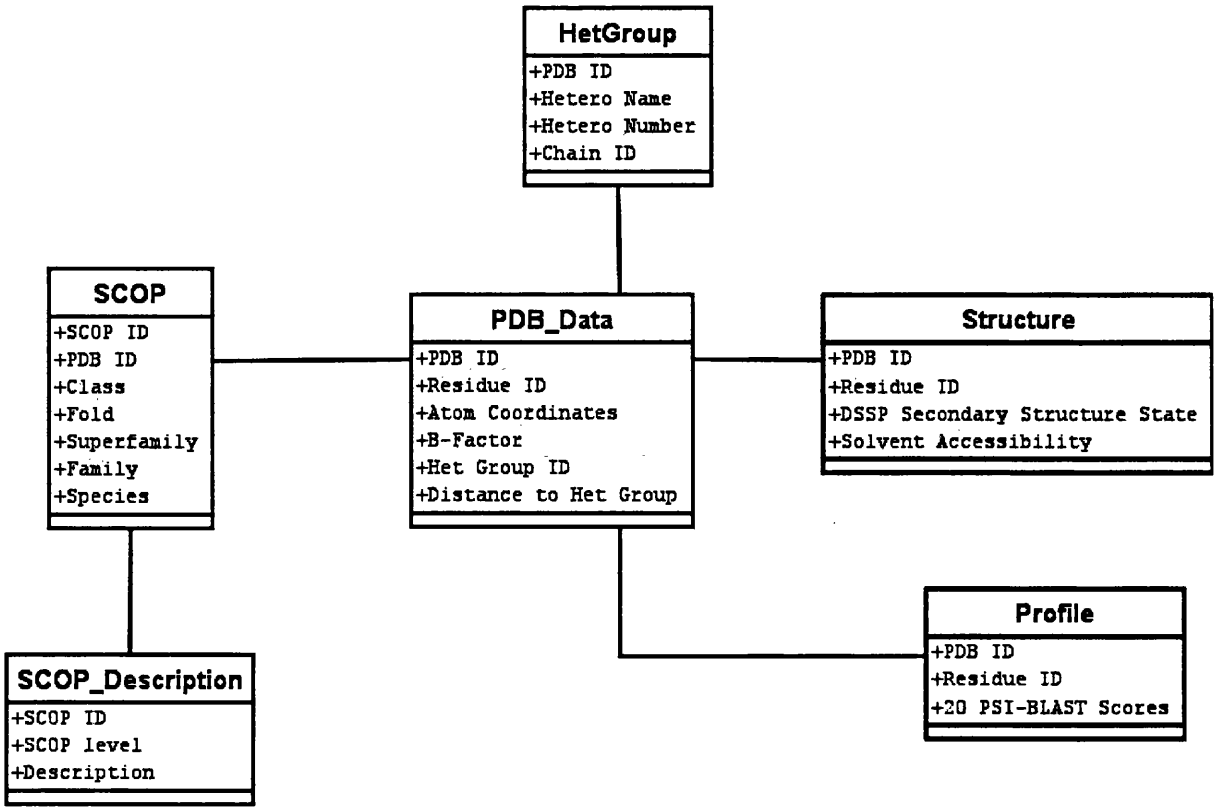


Figure 2.1: FuncSite Relational Database Schema

The Het group identifier as well as the number of Het groups present in the PDB file are stored.

2.2.3 PSI-BLAST PSSM

One of the most powerful features of PSI-BLAST (Altschul et al., 1997) is the ability to store and analyse the position specific scoring matrix (PSSM) generated at any given iteration. The information within PSSMs have been effectively used in a

variety of problem domains ranging from secondary structure prediction (McGuffin et al., 2000), fold recognition (Kelley et al., 2000) and disorder prediction (Ward et al., 2004).

It was decided to incorporate the information contained within the PSI-BLAST PSSMs into the FuncSite database. The profiles were generated for a list of protein chains containing the Het group of interest. This was done using the distributed PSI-BLAST tool developed by McGuffin et al. (2004b). The distributed system uses 50 dual processor nodes allowing PSI-BLAST score profiles to be generated in a practical time-scale for large datasets. Each protein chain was queried against the non-redundant database NRDB90 (Holm and Sander, 1998) which was compiled using a combination of SWISS-PROT, TrEMBL, GenBank, PIR, WormPep and PDB databases. Three iterations of PSI-BLAST were used with an E-value cut-off of 0.001.

2.2.4 Site Information

The *SiteLoad* program identifies residues within interacting distance of a target Het group and automatically extracts the corresponding matrix of 20 scores from the PSI-BLAST profile for interacting residues. In addition relative solvent accessibility and secondary structure state information is extracted from the DSSP (Kabsch and Sander, 1983) file of the corresponding PDB file.

Profile scores are loaded into the *Profile* table (from within *SiteLoad*) along with a residue identifier (derived by concatenating residue name and number). The structural data from DSSP for the site residues are stored in the *Structure* table whilst the $C\beta$ ($C\alpha$ for glycine) coordinates and associated temperature factors from the PDB file are inserted into the *PDBData* table.

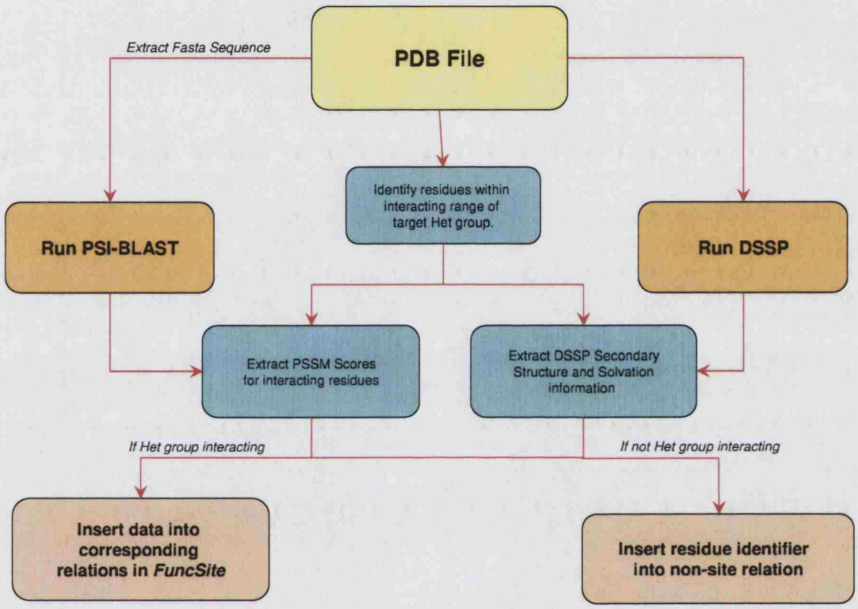


Figure 2.2: Flow diagram illustrating processing of sequence profile and structural data for FuncSite.

2.2.5 Querying Site Information

The Java class *SiteQuery* is used to automatically obtain the combined information from the profile and structure table for a given site type. The combined information is used to create a single table which is stored temporarily to allow faster querying for future analysis.

2.2.6 Structural Classification

The May 2003 version of SCOP (release 1.63) was parsed and loaded into the *SCOP* tables within FuncSite. Two tables are created, the first containing the SCOP identifier for each level of the SCOP classification scheme for every PDB. A second table is used to store the text based description at different levels for each SCOP entry.

2.3 Database Analysis

The aim of this section is to present various examples of analysis which can be performed using FuncSite. This is to highlight the type of information which can be extracted as well as providing a validation for the structural and sequence features captured by the system.

2.3.1 Prosthetic Groups Overview

FuncSite was queried to determine the overall number of different heterogeneous (Het) groups in the PDB. It is not uncommon for a given PDB structure to contain multiple instances of a given Het group, therefore only single occurrence of each unique Het group in a given PDB file were recorded for the current analysis. All Het groups observed at least 100 times are presented in Figure 2.3.

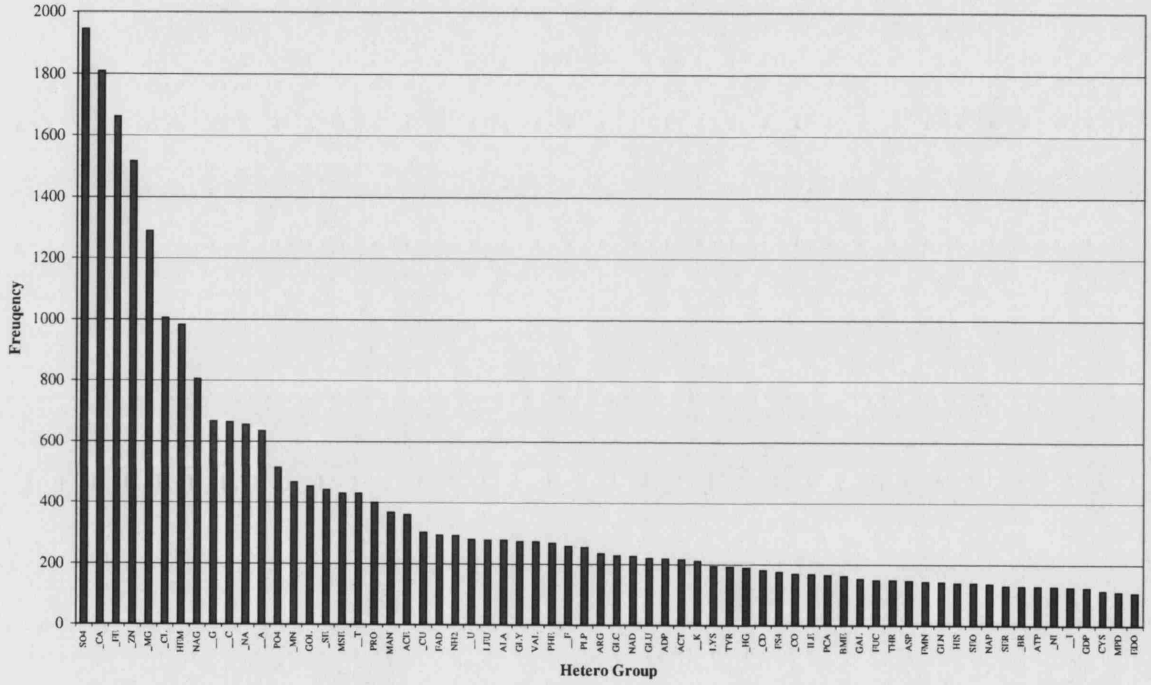


Figure 2.3: Frequency of occurrence of different Het groups in the PDB. For a given PDB file only a single occurrence of a given Het group is recorded. Figure includes all Het groups observed at least 100 times.

The initial broad queries of the database illustrate the dominance of ions within

protein structures contained in the PDB. Of the top twenty-five prosthetic groups present in the PDB seven are metals, Ca^{2+}, Fe^{3+}, Zn^{2+}, Mg^{2+}, Na^{+}, Mn^{2+} and Cu^{2+}.

One of the complications of the Het group data in the PDB is that it is often difficult to distinguish the presence of a Het group as being functionally significant or a non-specific artefact of the experimental technique used to determine the structure. This is strikingly illustrated by SO_4^{2-} which is the most abundant Het group in the PDB. However this particular observation is due to the wide spread use of ammonium sulphate as a buffer solution rather than of any biological significance.

Nonetheless, the presence of Het groups and their binding sites may provide interesting details relating to protein function and evolution. Metal binding sites in particular are known to be involved in a diverse set of biological functions, such as in the catalytic centres of enzymes as well as structural roles and signalling pathways (Chapter 1). The FuncSite resource was therefore used to analyse metal binding sites in the PDB to demonstrate the database and highlight trends within the captured structural and sequence data. The current analysis is focussed on features of proteins containing calcium ions (although analysis was also performed for other metals, see Appendix A).

2.3.2 Database Analysis of Metal Binding Sites

The properties surrounding metal binding regions were extracted from FuncSite and analysed. Several features were extracted, including the amino acid distribution, secondary structure content, residue conservation and solvation as well as SCOP superfamily distribution. The site information is compared to features extracted from non-site regions.

2.3.3 Amino Acid Propensity

FuncSite was queried to retrieve the amino acid distribution for residues $\leq$15Å from the metal ion of interest. The queries were performed for residues at varying distances from the position of the crystallised metal ion. These interaction were classified as short ($\leq$5Å), medium (5-7Å) and long-range (7-15Å) by measuring the Cα-metal ion distance. This allows site characteristics to be analysed in the context of distance from the metal. All residues beyond the 15Å cut-off are regarded as non-site residues. The amino acid distribution for residues surrounding Ca^{2+} ions is shown in Figure 2.4. The total number of residues within 15Å of a calcium ion was 32,234 compared to 126,021 residues in the non-site set.

The results clearly highlight that aspartate residues dominate the region directly surrounding the ion site comprising 41% of residues as compared to only 6.2% for the non-site data. Glycine is also more represented within 5Å of calcium, accounting

for 14.9% of residues. This is compared to 8% for non-sites. The significance for the remaining residue types, however, is less clear. Therefore a simple significance test was performed using the binomial probability function, pbinom, in the R open source statistical package (Ihaka and Gentleman, 1996). The following parameters were used for the function: number of trials, in this case the total number of residues within a given distance category, observed number of residue x, and the probability of observing residue x from the non-site data. The results highlighted that asparagine and glycine, within 5Å of calcium, are indeed statistically over-represented (p-values of 1.53×10^{-17} and 1.88×10^{-32} respectively). Additionally both glutamate and glycine, within 5-7Å of calcium, are also significantly more prevalent in calcium sites (p-values of 3.7×10^{-56} and 2.4×10^{-8} respectively)

The striking abundance of Asp residues in Ca^{2+} sites demonstrates the important contacts this residues makes when co-ordinating Ca^{2+} ions. These observations are consistent with the known chemical affinity of oxygen atoms to act as Ca^{2+} ligands which, in this instance, are located on the Asp and Glu side chains. Furthermore, the high proportion of glycine residues around calcium ions may be indicative of binding in loop regions connecting secondary structure elements, as for example observed in the EF-hand motif.

Interestingly, the higher proportion of Glu residues in the medium range, as compared to Asp, is probably due to glutamate having a longer side-chain, hence

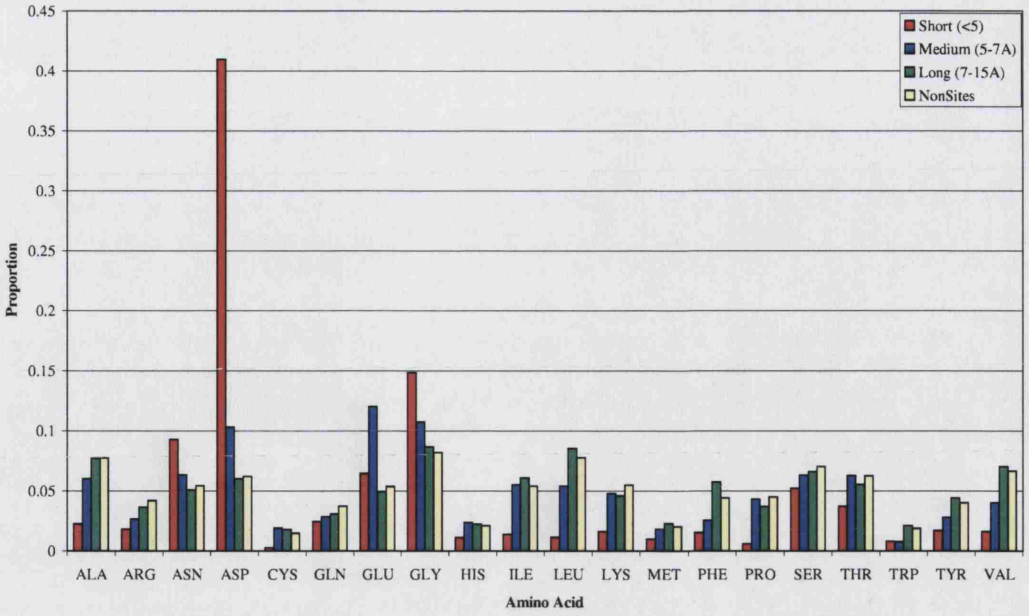


Figure 2.4: FuncSite database analysis: distribution of amino acid residues for protein chains containing calcium ions. Residues were categorised by distance from the calcium ion. The proportion represents the number of residues of a particular type in a given distance category over the total number of residues in that category.

the main chain atoms are placed further away from the metal ion.

2.3.4 Analysis of PSSM Scores

Residues involved in forming functional sites of proteins generally show a higher degree of conservation since they have evolved to perform the given function. Residues in non-site region, with the exception of residues required for maintaining fold integrity, are not expected to be under selective pressures and therefore will generally

be less conserved.

Many studies have used patterns in residue conservation in order to distinguish functionally important regions from other parts of the protein. In the current analysis the R statistical program Ihaka and Gentleman (1996) was used to generate several plots to visualise the trends of PSSM scores contained within *FuncSite*. Several examples are presented for the analysis of Ca^{2+} binding proteins.

Box Plots

The PSSM scores for Ca^{2+} sites and non-Ca^{2+} sites were used to construct box plots. Figure 2.5 clearly illustrates that although the extreme values are generally indistinguishable from the two populations, the majority of Asp site scores in Ca^{2+} site residues are shifted towards higher PSSM scores. This approach allows the distribution of all residue scores to be compared directly. Interestingly, other residues which are more prevalent around Ca^{2+} sites (Glu and Gly) are not as clearly delineated from the non-site scores as compared to the Asp scores.

Density Plots

Figure 2.6 shows the probability density plots for Asp PSSM scores for Ca^{2+} site and non-site residues. This clearly demonstrates the presence of two distinct populations of PSSM scores within residues located around calcium sites. The preference for

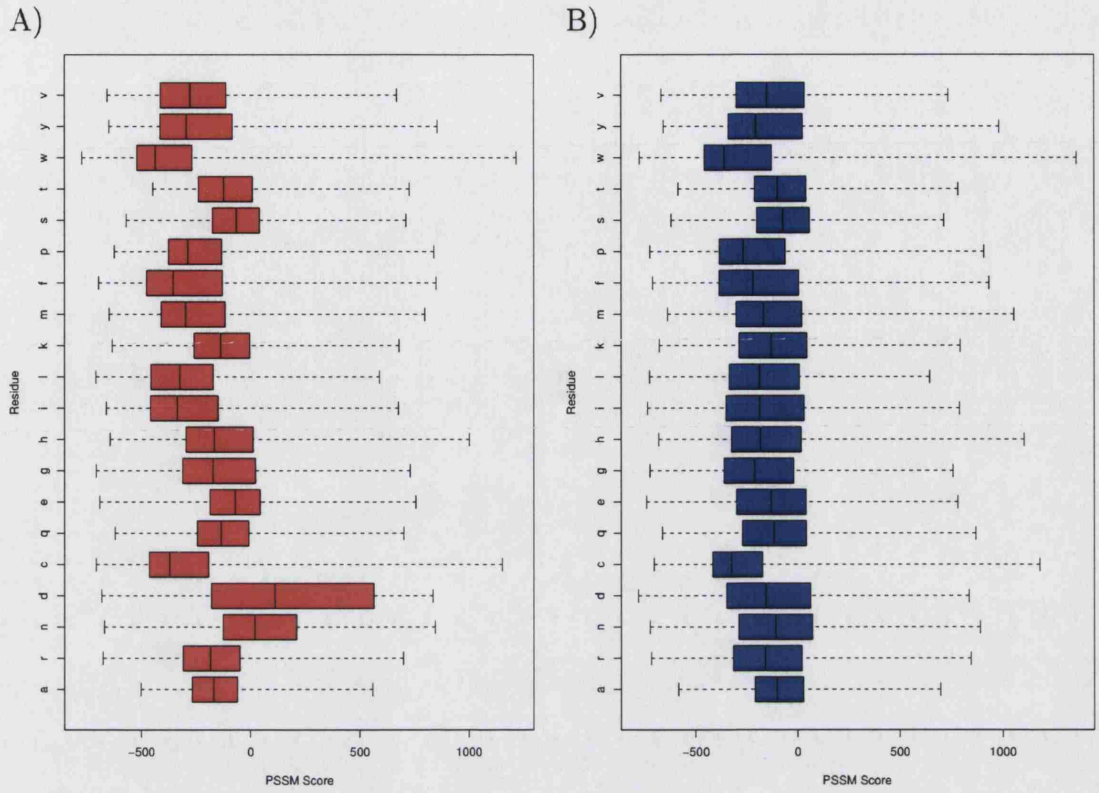


Figure 2.5: Box plot representation of the distribution of PSI-BLAST PSSM scores for: A) residues within 15Å (site) of calcium ions and B) residue outside this cut-off (non-site).

Asp residues is indicated by a second peak at higher PSSM scores and highlights the greater degree of conservation. Interestingly we observe that within the Ca^{2+} sites there is a large proportion of residues which do not show conservation and are indistinguishable from the non-site dataset. The advantage of this type of analysis is that it allows global trends within the PSSM values to be visualised and compared between two populations of data.

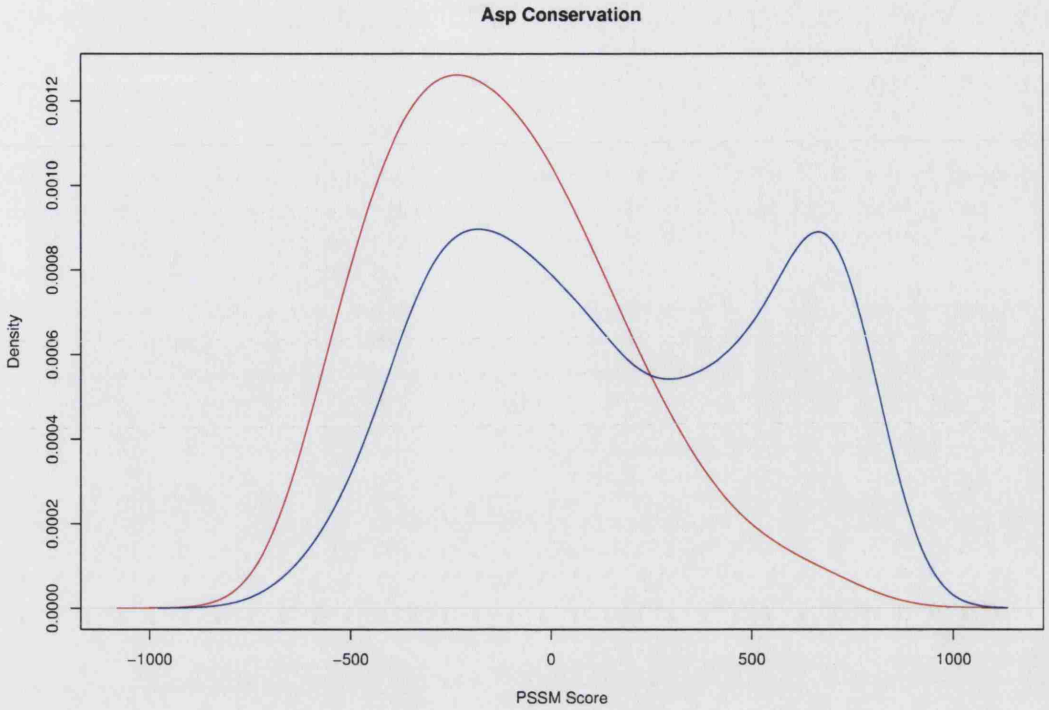


Figure 2.6: Distribution of aspartate PSSM scores for residues within calcium sites (blue) and non-site (red).

Contour Density Plots

Although the density plots demonstrate clear trends between PSSM scores of sites and non-sites for a single residue type, it does not show the PSSM score relationship between different residue types. Figure 2.7 illustrates the distribution of Asp scores in relation to Gly scores for Ca^{2+} site residues. This plot was generated by extracting the PSSM profile scores for Asp and Gly for all residues in Ca^{2+} sites. All residues are taken as we wish to analyse the degree of conservation for a given position in the

sequence (for either Asp or Gly) regardless of the actual residue at that position.

Interestingly, three distinct peaks are observed. The most prominent peak is observed for high Asp PSSM scores providing further evidence of the importance of these residues in calcium sites, this time in an evolutionary context. A peak, albeit less prominent, is also observed for high Gly scores. The last peak demonstrates a population of residues which are neither conserved for Asp or Gly residues. These cases are likely to be ligated by other residue types, for example glutamate. However, an alternative explanation could be the presence of calcium sites in the dataset which are not biologically significant and perhaps artefacts of crystallisation. Finally, we observe that highly conserved Asp scores are not observed when highly conserved Gly scores occur and vice versa. This illustrates the different physical properties of these two residues: Gly is small and often located in loop regions whilst the Asp side-chain is much larger and is associated with a strong negative charge. It is therefore not surprising that these residues are not simultaneously highly conserved for a given position.

The example that has been presented has focussed on two particular residue types for calcium sites. The analysis provides a clear indication of higher dimensional relationships within the PSSM scores.

A) B)

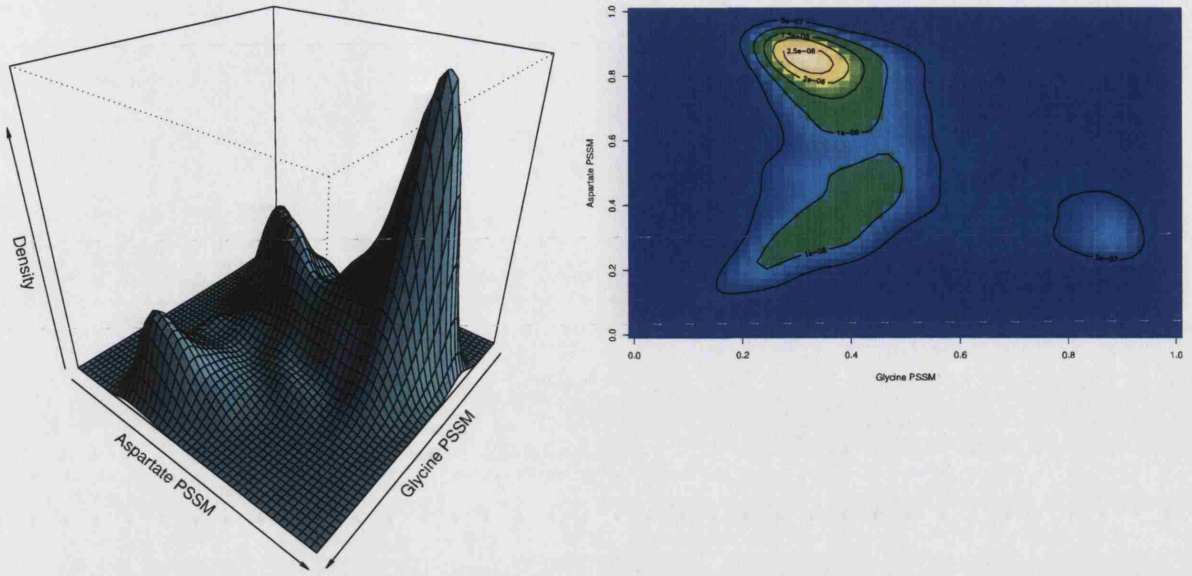


Figure 2.7: Two-dimensional plot for aspartate and glycine PSSM score distributions for residues located within calcium sites. A) Perspective plot and B) Contour plot.

2.3.5 Secondary Structure

The distribution of secondary structure states of residues surrounding metals ions and non-site regions were analysed. The six DSSP states were reduced to three states, H (α-helix, 3_{10}-helix), E (β-sheet, β-bridge) and C (turn, bend and other). The observations for the calcium binding proteins are shown in Table 2.1 (see Appendix A for other metal ions).

Overall there is a greater tendency of observing site residues in a coiled state. For the calcium set, 51.7% of residues which are within 5Å of a calcium ion are

	Distance from Metal			
DSSP SS	**≤5Å**	**5-7 Å**	**7-15 Å**	**>15Å(Non-Sites)**
Sheet	24.8%	24.8%	32.1%	24.4%
Helix	23.5%	27.2%	39%	30.4%
Coil	51.7%	48%	29%	45.2%

Table 2.1: Assignment of secondary structure (SS) from the DSSP program for calcium binding proteins. The DSSP SS states have been reduced to: Helix (α-helix, 3_{10}-helix), Sheet (β-sheet, β-bridge) or Coil (turn, bend and other).

observed to be coil. For the non-sites, 45.2% of residues are observed as coil.

2.3.6 Residue Solvation

The relative solvent accessibilities (RSA) calculated by DSSP were extracted from FuncSite for analysis. The RSA values were normalised by dividing the observed RSA by the maximum RSA for that residue, as described by Rost and Sander (1994). The non-parametric Wilcoxon rank sum test (within the R program) was used in order to determine the significance of the difference between solvation scores for calcium site residues as compared to non-sites. The null hypothesis for this test is that the normalised RSA scores for site residues are not significantly higher than residues from non-sites. Overall, the results showed a statistically significant tendency for higher RSA values for site residues (p-value of 5×10^{-3}) indicating the

majority of calcium site residues are more exposed.

2.3.7 Distribution of SCOP Codes

Queries were performed to identify the number of proteins belonging to unique super-families, as defined by SCOP, for the different metal datasets. Figure 2.8 illustrates the most populated super-families for the calcium containing PDB files. Proteins belonging to the EF-hand superfamily were the most dominant proteins containing calcium sites making up 43/551 (7.8%) of all chains in the dataset.

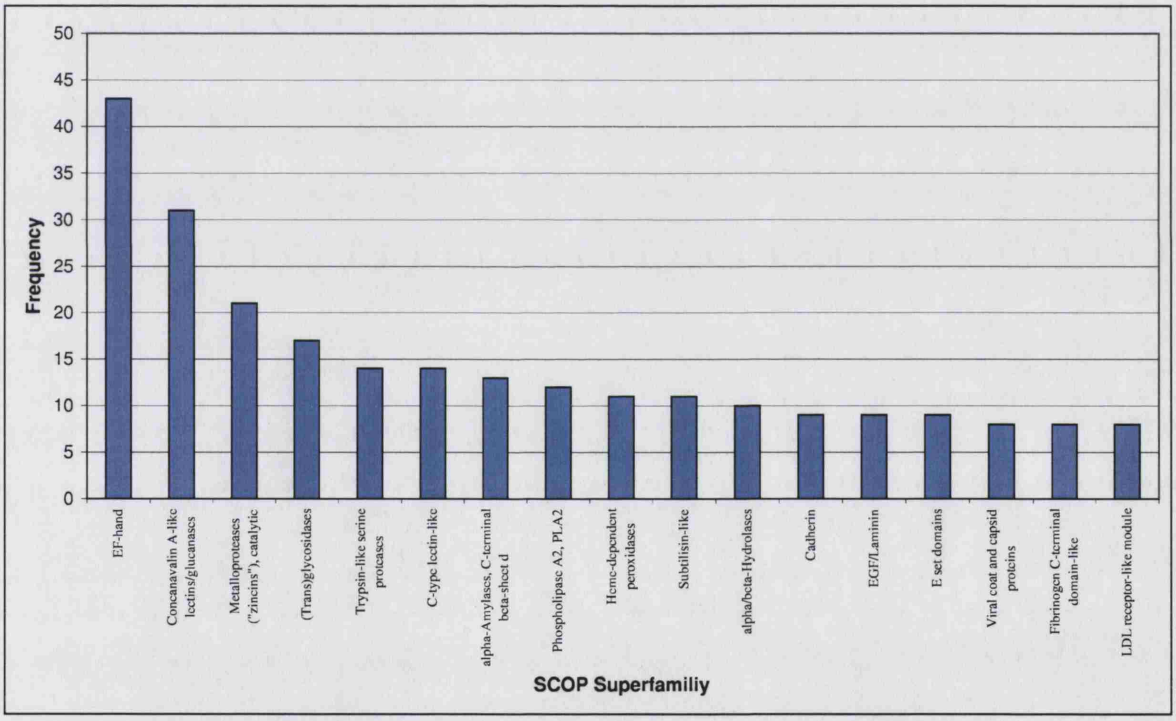


Figure 2.8: SCOP superfamily distribution of calcium containing proteins.

The superfamily analysis for the other metal containing protein chains highlighted several interesting findings. The most dominant zinc containing superfamilies were observed to be metalloproteases, immunoglobulins, glucocorticoid receptors (DNA binding domain) and zinc fingers. The magnesium containing chains were dominated by the p-loop containing nucleotide triphosphate hydrolases, representing 83 proteins whereas the second most populated superfamily for this set contained only 14 members (thiamin diphosphate-binding fold). Proteins containing copper were also dominated by one superfamily consisting of 47 members of the cupredoxin fold. The concanavalin A-like lectins/glucanases made up the largest group for the manganese set. The iron containing proteins showed a more even distribution of SCOP superfamilies. The periplasmic binding proteins made up the largest group of ten proteins whilst the superoxide dismutases and ferritin-like folds made up groups of eight proteins.

2.4 Discussion

The focus of this chapter has been the design and development of a new relational database model to combine information from sequence profiles and structural features of functionally important regions in protein structures. Several useful features of the system have been demonstrated for commonly occurring metal ion sites. Although the focus of the analyses has been on calcium binding proteins the database has been designed for the analysis of any site type. An important consideration for the current analysis has been to examine general trends rather than detailed high resolution features.

Metal co-ordination is known to be primarily directed through side-chain contacts. The database was therefore queried in order to determine if a distance based analysis of residues surrounding metal ions could validate the findings in the literature. For calcium sites, an increased proportion of aspartate, glutamate and glycine residues were observed as compared to non-calcium site regions. This is in accordance with the known affinity of calcium for oxygen atoms (Asp and Glu) and demonstrates such preferences are detectable using the current database.

Given the observed residue preferences, the PSI-BLAST PSSMs were analysed to identify global patterns of metal ion residue conservation. Several approaches were presented, using calcium sites as an example. The PSSM scores were observed to provide consistent results, highlighting greater tendency for conservation for Asp

residues around Ca^{2+} sites. Interestingly, analysis of the relationship between PSSM scores for Asp and Gly site residues highlighted a more complex underlying pattern of conservation.

Analysis of secondary structure derived from DSSP revealed an overall preference for coiled states in the vicinity of metal ions. For calcium this is likely to be a reflection of binding occurring in between secondary structure elements such as observed within the EF-hand motif which is the most prevalent calcium containing superfamily.

Analysis of relative solvent accessibilities (also from DSSP) indicated a statistically significant tendency for calcium site residues to be more exposed as compared to non-site residues. This is likely to demonstrate the fact that metal binding quite often occurs in regions between secondary structure states as well as toward the protein surface.

Another useful feature of FuncSite was demonstrated by incorporating the SCOP structural classification system. This allowed the distribution of different SCOP superfamilies to be determined for metal binding proteins. Such an analysis is useful at uncovering relationships between proteins containing a functional site of interest and higher level organisation of protein structures containing those sites.

Overall the database queries and analyses demonstrates significant features, known to be of importance in metal binding, to be effectively validated using the

database model. This highlights the aim of FuncSite: a framework for performing global as well as specific analyses of structural and sequence features. Although we have focused on metal site residues such an approach would be well suited to almost any functional region.

An important consideration, however, in site analysis is the effective definition of functionally relevant sites, certainly the presence of substrate, inhibitor or other prosthetic groups can provide strong indication of a site region. The preliminary identification of a functional site, and discrimination of these regions from bound sites which result from the crystallisation process is an important area.

The possibility of using a statistical based approach in order to classify site and non-site residues would aid in characterising such regions more accurately. Delineating the complex underlying features of such regions requires a more rigorous approach for defining sites and measuring similarity of features between sites.

In the next chapter we aim to use the features described in the current analysis to produce an automatic tool to locate and characterise metal binding regions in three-dimensional protein structures. One of the important aspects of the system is the focus on low resolution 'fuzzy' structural details in combination with the sequence profile information.

Chapter 3

Prediction of Metal Binding Residues using low-resolution features

3.1 Introduction

Understanding the relationship between protein structure and function has become one of the central goals in structural bioinformatics. The accurate prediction of biological function, on a genome-wide scale, promises wide ranging benefits in understanding complex biological processes as well as improving annotations to aid further research. This knowledge will be a key stepping stone in the development of techniques and pharmaceuticals to target disease genes and their products. However, in recent years it has become clear that effective and reliable tools are required to fully utilise the information provided by protein structure. This requirement is likely to grow as structural genomics initiatives rapidly solve protein structures for novel protein sequences.

The increased interest in predicting function from structure is highlighted in the literature with new methods being presented on a frequent basis. A common theme with many of these methods has been the dependence on highly accurate protein structures, as a consequence the analysis of low resolution protein structures has received relatively little attention.

The motivation behind the work presented in the current chapter has been the development of a new algorithm to allow information from structure and sequence to be combined allowing the automatic detection of functional sites. The approach has been developed for the prediction of metal binding residues using low resolution

features to allow site detection in low resolution models. In the following sections the motivation behind the study of protein-metal interactions will be discussed and an overview of other studies of metal binding presented.

3.1.1 Why Predict Protein-Metal Interactions?

It is surprising that, although metal ions are abundant in biology and are associated with diverse and important functions, relatively little bioinformatics research has been conducted in predicting structural aspects of metal binding. Estimates suggest that approximately one-third of all proteins require metal ions as cofactors for biological function (Holm et al., 1996). The biological metals include magnesium, calcium and the first transition series (vanadium, manganese, iron, cobalt, nickel, copper and zinc), molybdenum, tungsten, cadmium and mercury.

The accurate identification of metal binding regions can potentially aid the functional classification of a protein. Several enzymatic reactions require metal ions and the correct identification of metal type can be a useful discriminator of enzymatic function. Furthermore, metal binding sites may reveal important structural details and arrangements that can lead to improving the information available for a target protein.

The wide-ranging dependence of metal ion as prosthetic groups, from physiological to molecular level, have been studied in great detail. A discussion of metal

functions and several studies that have characterised metal ions in proteins follows.

3.1.2 Metal Functions

An extensive review of protein-metal chemistry, structure and function was presented by Holm et al. (1996). This review categorised metal function in proteins into five groups: i) catalytic ii) structural, iii) transport and storage, iv) electron transfer and v) dioxygen binding.

A particularly important catalytic role for metals is illustrated in the super-oxide dismutases (SOD); this family of enzymes is responsible for the efficient removal of highly reactive super-oxide free radical species generated during aerobic reactions. The catalytic mechanism, by which super-oxide dismutatses are able to sequester the highly damaging free radical oxygen species, fundamentally relies on the presence of metal ions. Commonly, zinc and copper are found in eukaryotic SOD's whereas manganese and iron are mainly found in prokaryotic SOD's (Holm et al., 1996). Within the active site the metal ion acts as a electron acceptor from super-oxide species resulting in hydrogen peroxide and molecular oxygen.

Metal ions are also frequently involved in the stabilisation and formation of protein structure. Thousands of different transcription factors have been shown to contain zinc ions (Branden and Tooze, 1998). The zinc finger motif provides an important example: this structural motif is commonly associated with DNA binding

function. In this configuration the zinc allows a stable loop region ('finger') to be formed that can be inserted into the major groove of DNA. The zinc ion is co-ordinated by two cysteine and two histidine residues which forms the foundation of the motif, a linker segment between one of the co-ordinating cysteine and histidine residues form the characteristic 'finger'.

3.1.3 The Metal Site Environment in Protein Structures

Understanding details of the molecular environment of metal ion sites in proteins is important to aid the elucidation of functional details. Karlin et al. (1997) presented a detailed article identifying similarities and differences between several commonly occurring metal sites. The study highlighted that metal ion sites consist of at least three layers: the metal core, the ligand group and the second shell. The most prominent protein ligands for Cu, Fe, Mn and Zn were reported as imidazole nitrogens from histidine, carboxylates (aspartate and glutamate), sulphur (cysteine, methioine), carbonyl oxygens and solvent.

Importantly, differences are observed depending on protein function and metal type. For instance, copper ions are never ligated by acidic residues and iron is predominantly ligated by multiple histidines. Manganese ions are primarily ligated by acidic residues whilst zinc ions participate in the most varied interaction including histidine, aspartate, glutamate, cysteine. In addition, zinc may be found to be co-

ordinated by tyrosine, asparagine, serine and threonine residues (Karlin et al., 1997).

3.1.4 Classification of Metal Binding Sites

Methods to effectively locate and classify metal binding regions in protein structures have been relatively scarce. Gregory et al. (1993) developed a high resolution system that measured the hydrophobicity contrast within the protein structures to locate the position of metal binding. The approach however was limited to only 28 protein structures available at that time.

(Wei and Altman, 2003) developed the FEATURE method to characterise well defined functional sites based on statistical descriptors derived from a set of site and non-site data. FEATURE also uses the exact placement of side chain atoms as well as incorporating secondary structural information but does not directly include conservation information for site residues. The method was applied to locate calcium binding sites in a set of model structures (Wei et al., 1999) and was shown to require high resolution placement of atoms specific to the site region. Therefore the major drawback of the FEATURE approach is the reliance of highly accurate protein structures, and are therefore unsuitable for low-resolution protein models.

Degtyarenko (2000) presented the concept of bioinorganic motifs (BIM) to aid the functional classification of metalloproteins. A BIM is defined by the metal ion and its first coordination shell ligands and can allow more effective chemical comparisons

between metal binding regions. The metalloprotein database (MDB), at the Scripps institute (Castagnetto et al., 2002), provides BIM like information on protein-metal binding sites from structures in the PDB. Analytical tools are available to examine trends in the data contained in the MDB.

3.1.5 Chapter Overview

The focus of this chapter is the development of a novel approach using artificial neural networks (ANN) to predict six commonly occurring metal ion sites ($Ca^{2+}, Cu^{2+}, Fe^{3+}, Mg^{2+}, Mn^{2+}$ and Zn^{2+}). The method is designed to identify residues forming the metal binding site in superfamilies by combining sequence profile and structural information. Although the classification system is developed and benchmarked for metal binding sites, in principle, there is no reason the approach cannot be extended to other types of sites.

The motivation of the study has been the development of functional site predictors where only moderate quality structural information is available. Metal binding site predictions are assessed for a set of newly released structures from LiveBench (Rychlewski et al., 2003) illustrating effective site detection in novel proteins. A metal binding prediction for a LiveBench target is validated using the literature. In addition, we report a putative metal binding site predicted in a structural genomics target with unknown function. Site detection is demonstrated to be effective in low-

resolution predicted structures generated by the GenTHREADER fold recognition system. Finally, the classifiers were made publicly available through a web server.

3.2 Materials and Methods

3.2.1 Datasets

The training set was constructed by taking all protein chains interacting with the specified metal ions from the PDB and clustering at a 25% sequence identity, this resulted in 1018 sequence clusters. For the purposes of cross-validation these chains were then grouped into 364 distinct SCOP super-families. The number of PDB chains, super-families and metal sites in the dataset is summarized in Table 3.1.

Metal Ion Type	**No. of PDBs** Chains	**No. of SCOP** Super Families	**No. of Metal** Ions
Zn^{2+}	512	190	803
Ca^{2+}	443	128	819
Mg^{2+}	349	124	470
Mn^{2+}	168	49	253
Cu^{2+}	86	11	110
Fe^{3+}	70	18	83

Table 3.1: Summary of dataset used to develop the MetSite method.

3.2.2 Site Features and Definitions

The PSI-BLAST score matrices were derived by performing three iterations of PSI-BLAST against a non-redundant database (NRDB90) for all the unique chains across

all datasets. Any residue with main chain $C\beta$ ($C\alpha$ for glycine) atoms within 7Å of a target metal ion were defined as a site seed residue. The N closest neighbouring residues to the seeds were marked as seed neighbours. Figure 3.1 illustrates the site encoding.

For each of the marked residues, several features were calculated. These included the 20 scores taken from the PSI-BLAST PSSM, secondary structure state (reduced to Helix, Sheet or Coil), solvent accessibility from DSSP (Kabsch and Sander, 1983) and finally the inter-atomic distances between the $C\beta$ ($C\alpha$ for glycine) atoms for the site residues. Thus for each site, consisting of N residue (seed + N-1 neighbours), we defined 20N PSI-BLAST profile scores, 3N secondary structure states, N solvent accessibility scores and an inter-atomic distance matrix between the N residues ($\frac{N(N-1)}{2}$).

The metal ion environment that forms the second coordination shell as well as regions further away from the ligand residues are known to play a crucial role in site selectivity, however in the context of the site description used here, increasing the number of residues used to compose a site pattern results in over-fitting problems during training. We therefore restricted N to 10 which equates to 285 features and was shown to allow effective generalization properties.

Classification was also performed using PSSM scores for residues proximal in protein sequence. As for the site based method above all residues within 7Å of a

metal ion are marked as seed residues, however the neighbouring residues are those local in sequence.

3.2.3 Pre-processing Site Data

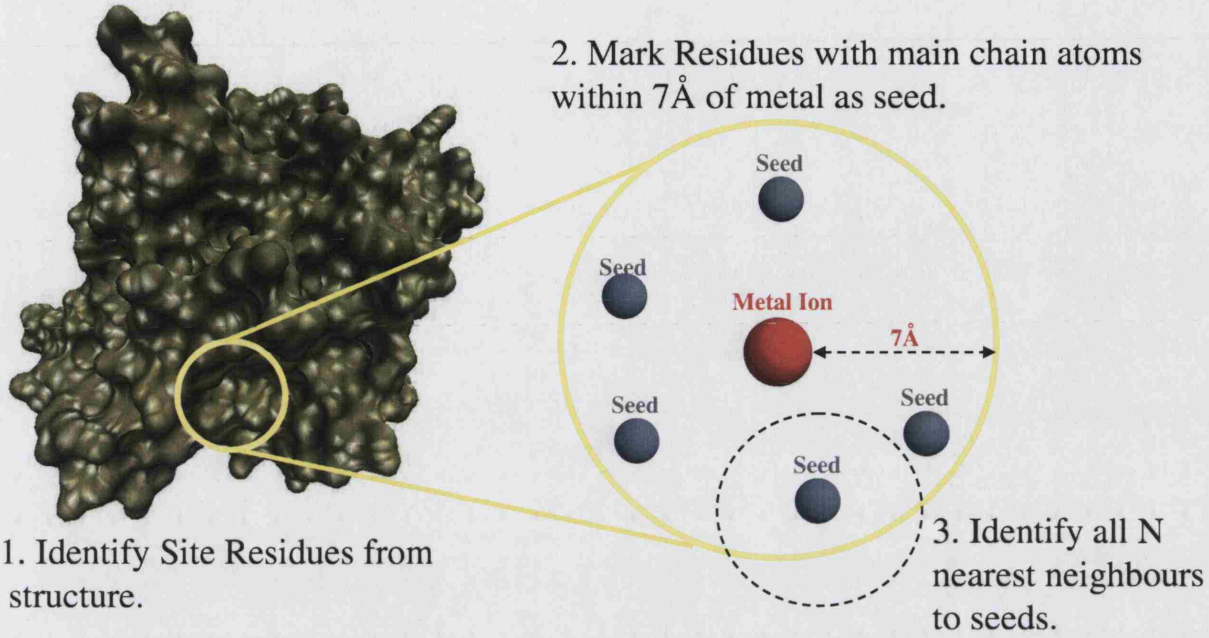
Prior to neural network training all values were re-scaled in the range [0,1] by using the standard sigmoid logistic function:

$$y = \frac{1}{1 + \exp^{(-ax)}} \tag{3.1}$$

where x is the raw input value, a is an arbitrary constant and y is the rescaled value.

3.2.4 Neural Network Training

For each metal type classifier a three-layer single output feed-forward neural network was created and trained using the neural network toolbox in Matlab (The MathWorks Inc., MA, USA). The number of nodes in the hidden layer is an important factor in preventing over-fitting on the training data especially in cases where limited training examples are available. It was found that 25 nodes in the hidden layer, trained using resilient back-propagation Riedmiller and Braun (1993) and early stopping resulted in good generalisation on the testing sets. For each training run 10% of the training data was used as a validation set to evaluate the performance of the network during training and prevent over-fitting. A more detailed discussion of



		A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
1	C	-1	-4	-3	-4	10	-4	-4	-3	-4	-2	-2	-4	-2	-3	-4	-1	-1	-3	-3	-1
2	G	0	-3	-1	-2	-3	-2	-3	6	-3	-5	-4	-2	-3	-4	-3	-1	-2	-3	-4	-4
3	V	2	-1	-2	-3	-2	-1	-2	-2	-3	1	-1	2	-1	-2	-2	-1	-1	-4	-2	3
4	P	-1	-3	-3	-2	-4	-2	-2	-3	-3	-3	-4	-2	-3	-4	8	-1	-2	-4	-4	-3
5	A	2	-1	-1	-1	-1	3	0	-1	-1	2	-1	-1	0	-2	-1	0	-1	-2	-1	0
6	I	-1	-2	-2	-2	-1	-2	-2	-2	-2	3	0	-2	0	4	-2	-2	-1	-1	0	3
7	Q	-2	3	-1	-1	-4	5	3	-3	0	-4	-3	1	-2	-4	-2	-1	-1	-3	-2	-3
8	P	-1	-2	-2	-1	-4	-1	2	-3	-2	-4	-4	-1	-3	-4	7	-1	-2	-4	-3	-3
9	V	-2	-1	2	-2	-3	-1	-2	-3	-2	-1	-2	2	-2	3	4	-1	-1	-3	-1	1
10	L	-1	-1	-2	-3	-2	-1	-2	-3	-2	-1	3	2	0	4	-3	1	-1	-2	0	-1
11	S	-1	2	-1	-1	5	0	2	-2	-2	-3	-3	0	-2	-3	-2	3	0	-3	-3	-2
12	G	1	-2	-2	-1	-3	-1	2	2	-2	-3	-3	-1	-3	-4	5	-1	-2	-4	-3	-3

4. Extract PSSM scores, DSSP assignments and calculate DM for all marked residues.

Figure 3.1: Site encoding scheme. DM = distance matrix. All residues within 7Å of a metal ion are labelled as metal interacting (seeds). A site pattern is derived by extracting sequence and structural information for residues labelled as seeds or neighbours to a seed.

neural networks and training is provided in Appendix B.

Transfer Functions

The logarithmic sigmoid and hyperbolic tangent transfer functions, defined below, were used between the input-hidden and hidden-output layers respectivley.

$$y = \frac{1}{1 + \exp^{(-x)}} \tag{3.2}$$

$$y' = \frac{2}{1 + \exp^{(-2x)}} - 1 \tag{3.3}$$

where x is the raw input value at each layer, and y and y' are the rescaled values for the first and second layers respectivley.

3.2.5 Cross validation

We screened all PDB chains within the dataset for each metal ion in turn. All residues outside the interacting range of a given metal ion were taken as negative sites. The 1018 sequence clusters were grouped at the SCOP superfamily level resulting in 364 clusters. We randomly split these SCOP clusters into five groups and carried out five-fold cross validation. This rigorous validation approach ensures that no two site patterns between training and test sets had any similarity at the superfamily level therefore mimicking site detection in new superfamilies. This allows the

generalization characteristics of each Metal Site (MetSite)classifier to be assessed in an extremely robust manner.

3.2.6 Assessing Performance

The results from the complete cross-validation test for each of the metal site datasets were pooled together to calculate prediction accuracy, defined as the total number of true positive and true negatives over all patterns (Q2). We also assess TPR and FPR as defined below. The non-parametric Wilcoxon statistic, representing the area under the ROC curve, is a robust measure for comparing classifiers and was calculated to determine the significance between classification of the feature sub-set.

$$\text{TPR} = \frac{(TP)}{(TP + FN)} \tag{3.4}$$

$$\text{FPR} = \frac{(FP)}{(FP + TN)} \tag{3.5}$$

For site based predictions we sum network scores for residues above a given threshold occurring within a 7Å region in the protein structure, site based sensitivity and selectivity was calculated as follows:

$$\text{Site Selectivity} = \frac{(TP)}{(TP + FP)} \tag{3.6}$$

where T = True, F = False, P = Positive, N = Negative and R = Rate.

3.2.7 Estimating Confidence

In order to estimate a confidence value for the predictions we analysed the cross validated network scores for the positive and negative sites for each metal ion (collaborative work, Bryson, K). Using the open source R package (Ihaka and Gentleman, 1996), we calculated the log_{10} of the ratio of positive cases to negative cases over 20 equal sized bins along the score range from 0.0 to 1.0. These were plotted, together with their standard errors, and minimum order polynomial fits were determined. In all cases except copper, 5th order polynomials gave satisfactory fits to the log ratios when taking their standard errors into account. For copper, a 9th order polynomial was required to give a satifactory fit. The resulting equations were used to convert network score outputs into log likelihood ratios for each metal type.

3.2.8 Visualization of Metal Site Predictions

The program MetPred was developed to produce a PDB formatted file where the temperature factor column is replaced by the neural network output score. MetPred takes in as inputs the weights obtained from the fully trained classifier and the PDB file to be queried. The results can be viewed in any standard molecular graphics-viewing program to highlight spatial clusters indicating likely metal ion interacting residues. All structures in this study were prepared and rendered using the VMD molecular graphics program (Humphrey et al., 1996).

3.3 Results

3.3.1 Feature Analysis

In order to determine the key features which allow effective discrimination of metal sites from non-metal sites, five-fold cross validation experiments were performed using only a subset of the site features (see Materials and Methods). During benchmarking we ensure that no two protein chains occur within the same SCOP super family between training and testing sets.

The classification results for individual feature sub-sets are illustrated in the form of receiver operating characteristic (ROC) plots in Figure 3.2. The plots indicate that structural information alone is not sufficient for sensitive classification, although it tends to marginally improve on the classification results using only PSSM scores of site residues. This highlights the important contribution of residue conservation in metal binding residues and indicates a clear functional relevance.

On average, inclusion of PSSM scores together with secondary structure, site residue distances and solvent accessibility resulted in a 94.5% Q2 accuracy with a True Positive Rate (TPR) of 39.2% at a 5% FPR threshold (Materials and Methods). Classification was marginally worse when training was performed using only the PSSM scores of site residues (TPR of 36.2%). For comparison training was also performed using PSSM scores for residues local in primary sequence as opposed to

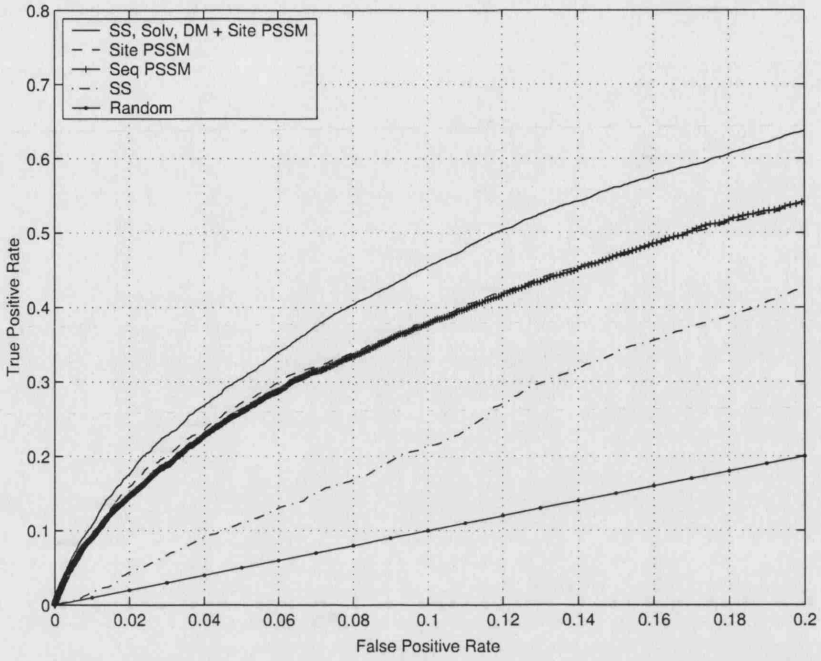
local in structure, this resulted in a TPR of 30.5%. Finally classification using only the secondary structure assignments gave a average TPR of only 13.7%. The average Wilcoxon statistic, which relates to the area under the ROC curve, between the different classifiers was calculated to be 81.1 (where 100 represents perfect classification).

The cross-validated neural network classification results were compared to predictions derived from a simple baseline prediction method using PSI-BLAST PSSM log likelihood scores. Metal-binding prediction was performed specifically for those residue types known to be more frequently occurring in the target metal sites. For example, in the case of calcium sites, the PSSM log likelihoods for Asp and Glu residues were extracted. Overall, we found that only 6.2% of metal-binding residues were predicted correctly using this approach at a 5% FPR threshold. Table 3.2 shows the overall cross-validation classification results for the naïve baseline and the full method on all sequence clusters.

Metal Ion	Q2 Accuracy	TPR	Wilcoxon	Naïve TPR(%)
Ca^{2+}	93.9%	30.4%	79.9	5.0
Cu^{2+}	94.9%	36.2%	85.6	2.7
Fe^{3+}	94.9%	48.8%	84.0	8.9
Mg^{2+}	94.2%	32.4%	73.8	6.8
Mn^{2+}	94.7%	38.8%	80.8	8.0
Zn^{2+}	94.6%	47.8%	82.2	6.0

Table 3.2: Cross-Validation classification results at 5% False Positive Rate (FPR). The True Positive Rate (TPR), Q2 accuracy and Wilcoxon statistic are defined in the Methods. The naïve TPR represents a baseline using residue conservation score.

a)



b)

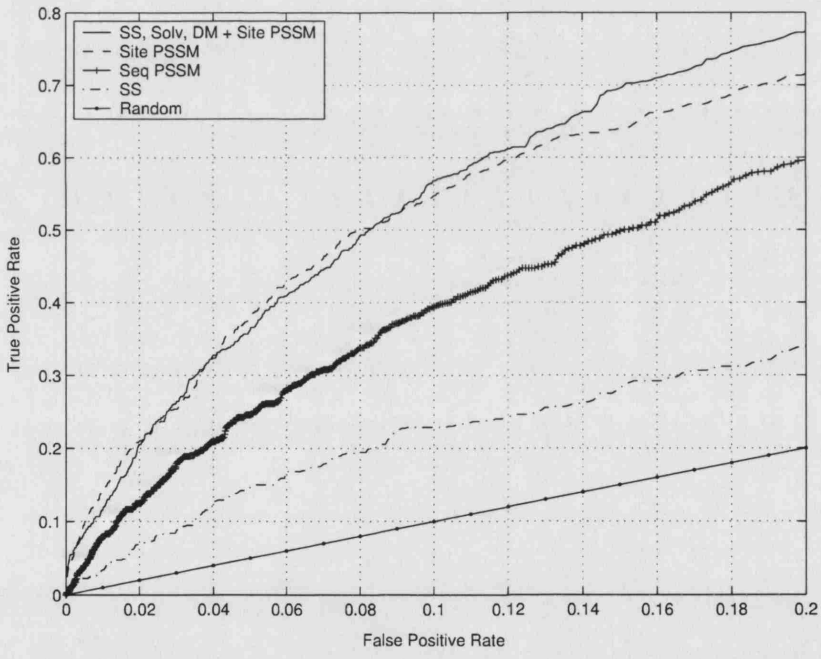
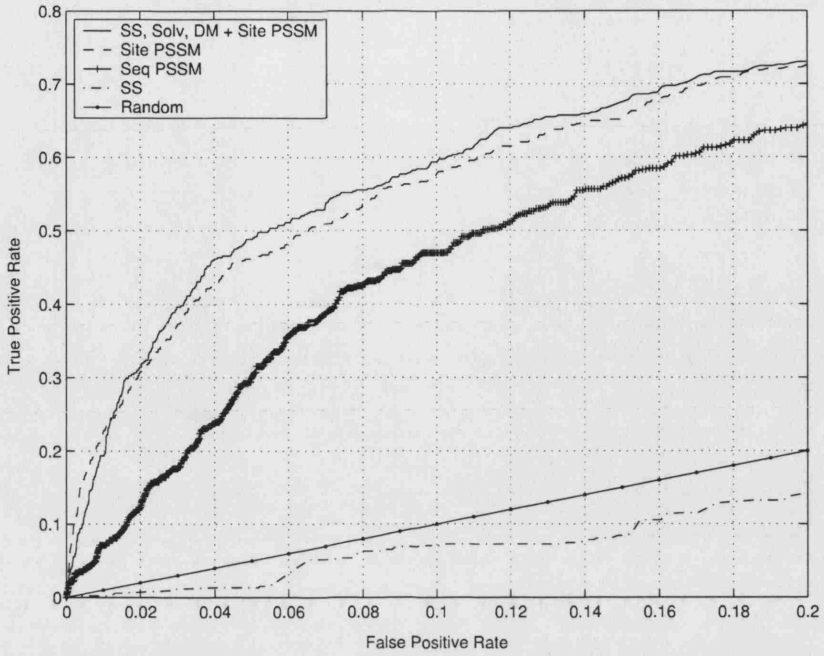


Figure 3.2 continued

c)



d)

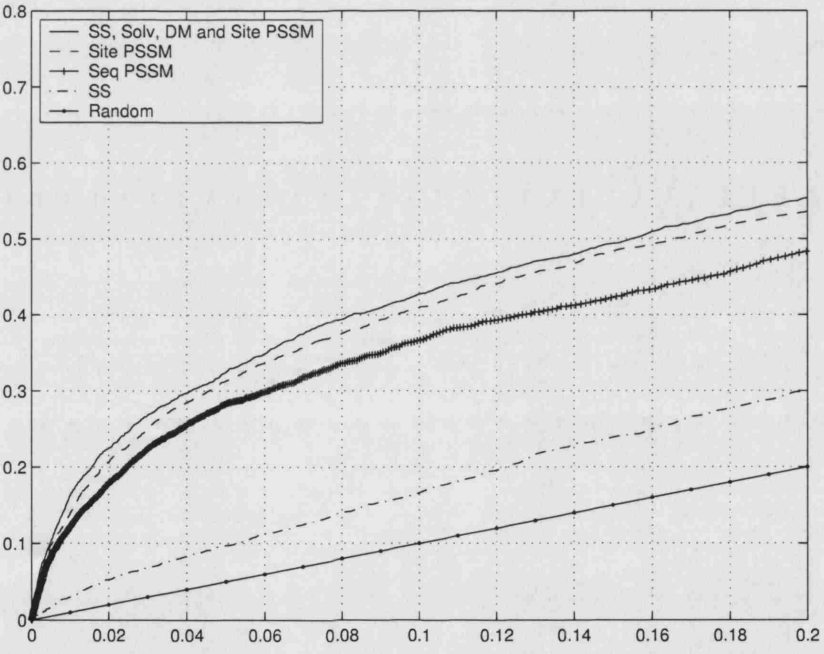
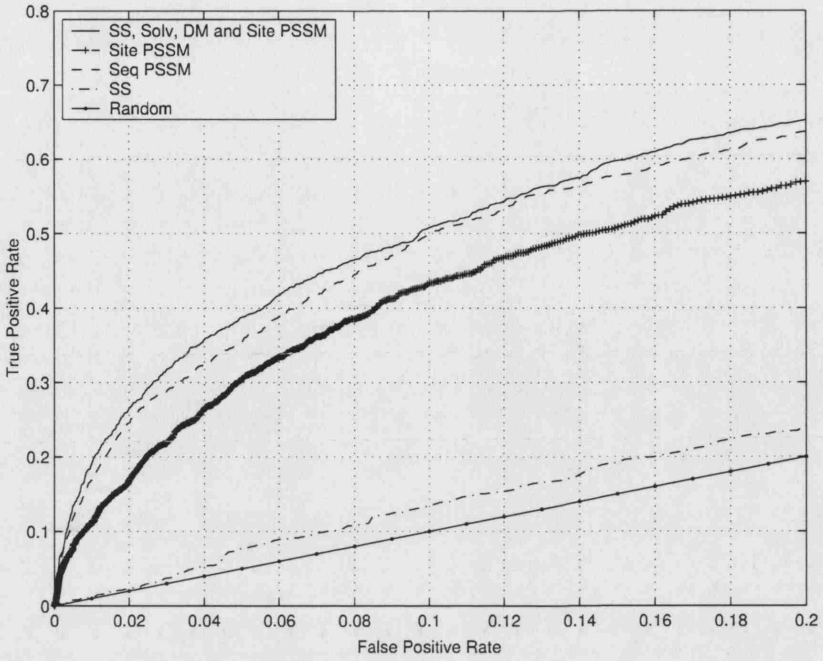


Figure 3.2 continued

e)



f)

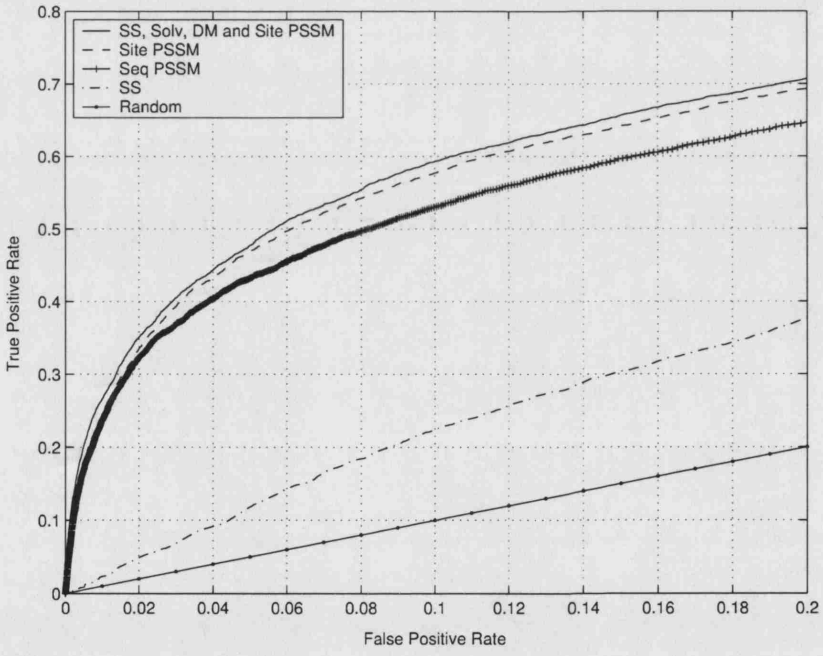


Figure 3.2: ROC plot analysis for a) Ca^{2+}, b) Cu^{2+}, c) Fe^{3+}, d) Mg^{2+}, e) Mn^{2+} and f) Zn^{2+} classification. Features investigated; secondary structure (SS), solvent accessibility (Solv), Position Specific Scoring Matrix (PSSM) and distance matrix (DM) defined as distances between site Cβ atoms (Cα for glycine). Classification was assessed for residues local in the 3D structure (Site) or local in primary sequence (Seq PSSM).

3.3.2 Site Based Detection

The results indicate that a large proportion of patterns, defined as sites under the above definition, are not correctly retrieved during cross validation under the allowable 5% false positive threshold. This is most probably due to the fact that there are many more site patterns than actual sites due to the encoding scheme employed here (Materials and Methods). For example, the 405 Calcium sites produce 3529 site patterns (i.e every site residue generates a site pattern). The MetSite method was therefore assessed by the ability to predict unique site regions. This was achieved by calculating a site score, determined by summing the neural network outputs for individual residues. All single residue hits above a specified threshold, and within a 7Å radius of a given target residue, were grouped giving a single score for that region of the protein. This post-processing has the effect of eliminating single residue false positive hits by attributing higher scores to regions containing several hits. The clustered site score approach resulted in the correct prediction of 60% of all metal sites in all superfamilies within the top ranking MetSite predictions.

3.3.3 Site Prediction in SCOP Super Families

The cross-validated classification performance was also investigated for each of the most populated SCOP superfamily clusters. The top ranking site predictions for these over-represented superfamily members is presented in Table 3.3 and clearly

indicates significantly better metal site predictions. Within the calcium containing protein chains the EF-hand like domains made up the most prevalent cluster consisting of 71 unique sites, of these 61 (85.9%) were correctly predicted with a selectivity of 73.5% (see Materials and Methods). Similarly site detection was much more accurate for all structures in the over-represented families where metal binding shows a clear functional relevance. Overall site sensitivity of metal sites for these clusters was 85% with a selectivity of 39%. Given that the neural network is trained such that no two proteins in training/testing fall within the same SCOP superfamily these results indicate MetSite to have effectively generalized site characteristics.

3.3.4 Confidence and Distinction between Metal Sites

It is essential that results have accurate confidence values assigned to them to permit the statistical significance of any finding to be assessed. Also, in the case where several networks produce high scores, we need to predict the most likely metal binding site. In order to accomplish this, we determine the log likelihood ratio for correct prediction against network score for each of the networks (Materials and Methods).

The log likelihood ratio of correct prediction against network output score is given in Figure 3.3. This allows us to rank the confidence of different prediction methods as a function of network score. For instance, at a network score of 0.9, zinc

SCOP Super Family	Super Family Representative	Total Sites	Site Sensitivity (%)	Site Selectivity (%)
Calcium				
EF Hand	1alvB	71	85.9	73.5
Phospholipase A2	1g4iA	6	100	40.0
C-Type Lectin	1byfA	6	75.0	54.5
Concanavalin A	1ajkA	15	60.0	23
Zinc				
C2H2 and C2HC zinc fingers	1f2iK	24	72.7	57.1
Metalloproteases ("zincins")	1ast0	17	58.8	26.3
Glucocorticoid	1a6yB	20	80	55.2
NAD(P)-binding Rossmann-fold	1e3lA	6	83.3	38.5
Zn-dependent exopeptidases	1cg2B	12	91.6	40.7
Magnesium				
P-Loop Hydrolase	1a820	35	88.6	32.0
ATPase Domain	1byqA	5	90.0	45.5
Phosphoenolpyruvate/pyruvate	1dxeA	4	100	27.7
Protein Kinase	1blxA	9	88.9	30.8
Copper				
Cupredoxin	1a4aA	32	62.5	44.4
Manganese				
Fe/Mn SOD	1gv3A	1	100	50
Iron				
Ferritin	1b71A	15	100	32.0
Rubredoxin	1b13A	4	100	33.3
Fe/Mn SOD	1gv3A	1	100	33
Clavaminate synthase-like	1bk00	3	66.7	10

Table 3.3: Site based predictions in over represented SCOP superfamilies. Site sensitivity and selectivity relate to the prediction of uniqe sites in the protein structure as described in Methods.

prediction is the most confident followed by magnesium, manganese, iron, calcium and finally copper. This permits a direct comparisons to be made between the different classifiers,for example more reliability should be attributed to a score of 0.7 for an iron prediction than a score of 0.9 for a calcium prediction.

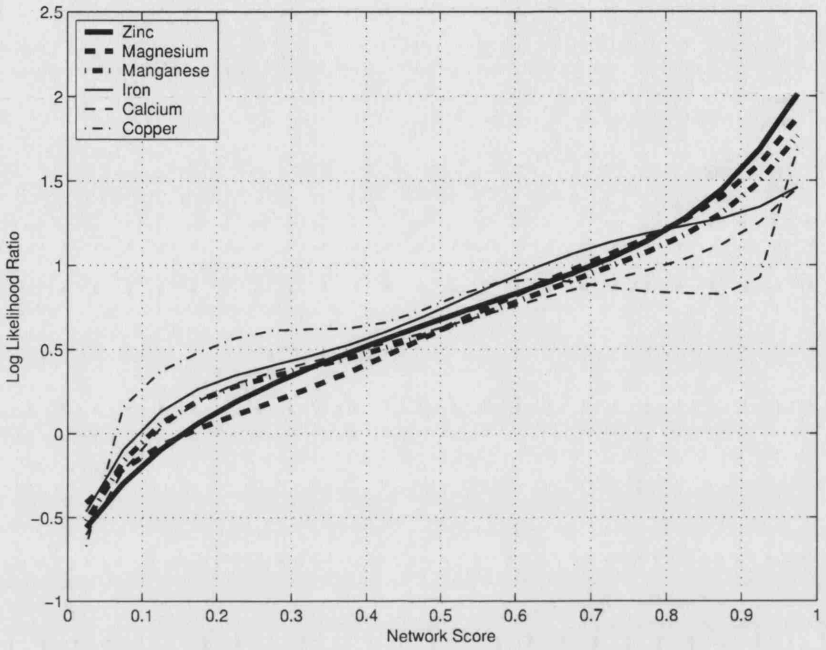


Figure 3.3: Likelihood ratio plots for the different metal site classifiers used to assess the statistical significance of the raw neural network outputs.

3.3.5 Site Prediction in LiveBench Targets

The LiveBench project (Rychlewski et al., 2003) is a continuous structure prediction assessment for newly released structures, including targets from the various structural genomics projects. These target structures are of particular interest as

they show no significant sequence similarity to any other known protein structures. MetSite was used to scan 172 protein chains from LiveBench-8, of these 24 chains contained occurrences of target metal ions in the crystal structure. The top ranking MetSite predictions correctly identified the true metal binding region in 19/24 (71.2%) of the crystal structures.

3.3.6 Identification of POP2 Metal Binding Site

The RNase domain of the yeast POP2 protein (1uocA) was predicted to bind Mn^{2+} with high confidence in a site region centered around Ser44. The predicted site was devoid of any prosthetic group although the protein did contain several calcium ions at different site regions. Inspection of the literature revealed the active site Ser44 of this protein is in fact involved in Mn/Mg binding (Thore et al., 2003) and makes up part of the active site region. The authors speculate that POP2 binds only a single metal ion instead of two metal ions observed for the DNases resulting in a different reaction mechanism. This is consistent with the MetSite predictions as only a single strong hit was observed, even though no metal is present in the active site of the crystal structure.

3.3.7 Modeling of LiveBench Targets

Using the mGenTHREADER fold recognition method (Jones, 1999; McGuffin and Jones, 2003) we assessed the ability of MetSite to locate metal binding regions in

structural models. It is possible to obtain good quality models using fold recognition techniques such as mGenTHREADER even when there is scant sequence identity between target and template. Models derived from mGenTHREADER alignments do not contain coordinates of side-chain atoms but rather the predicted location of the target protein's backbone, nevertheless such models are suitable for scanning with MetSite which requires only the approximate locations of residues.

All proteins within the LiveBench dataset were screened against a non-redundant fold library using the distributed version of mGenTHREADER (McGuffin et al., 2004a,b). In order to remove trivial cases where binding sites could be deduced from simple homology searches, the analysis was focused on target/template pairs that showed <30% sequence identity. In addition all template structures did not contain any obvious metal binding sites. The secondary structure of each target protein was predicted using PSIPRED (McGuffin et al., 2000).

The mGenTHREADER predictions of the 24 metal containing LiveBench targets produced 15 models with MaxSub (Siew et al., 2000) scores >0 (where 100 indicates a perfect structural prediction) in the top ranking hits (Table 3.4).

LiveBench Target	Template	SeqID (%)	MaxSub	mGT E-value
1o4tA	1lr5A0	15.0	66	0.05
1m3uA	1mumA0	13.3	42	2×10^{-4}
1oy0A	1mumA0	13.2	41	4×10^{-4}
1iujA	1lq9A0	16.0	54	0.04
1qvjA	1kt9A0	21.2	20	0.05
1mzbA	1qbjA0	20.0	38	0.16
1rifA	1qvaA0	15.5	36	0.01
1uf3A	1nnwA0	13.2	38	3×10^{-3}
1ei6A	1fsu00	9.1	25	3×10^{-4}
1uocA	1fxxA0	10.4	37	0.01
1qv9A	1fxxA0	11.7	16	0.03
1jwqA	1di0A0	10.1	29	0.08
1o4zA	1dypA0	15.9	42	4×10^{-3}
1o4yA	1dypA0	15.9	39	3×10^{-3}
1uetA	1fa0B0	10.1	25	2×10^{-4}

Table 3.4: Top ranking mGenTHREADER predictions for metal containing LiveBench-8 targets. The E-Value corresponds to mGenTHREADER confidence and MaSub relates to model quality. The bold targets represent cases for which metal sites were correctly predicted in the model structure.

3.3.8 Predicting Sites in Fold Recognition Models

Of the 15 metal containing LiveBench model structures MetSite correctly identified the metal binding region in 8 targets (53%) within the top site predictions (Table 3.4). Figure 3.4 illustrates the MetSite predictions for the native and model structures of zinc containing phosphonoacetate hydrolase (1ei6A) and the catalytic domain of N-acetylmuramoyal-L-alanine amidase (1jwqA). An InterPro search of these protein sequences did not reveal any obvious metal binding motifs.

The top ranking MetSite predictions indicate that all residues chelating the zinc

ion in both proteins have been correctly identified and cluster around the actual metal site with no false positive hits. The top ranking mGenTHREADER result for 1ei6A predicts a hit with the template structure human arylsulfatase (1fsu0) with an E-value of 3×10^{-4}, although the sequence identity between these proteins is only 9.1%. Figure 3.4c illustrates the best MetSite predictions for the 1ei6A model structure, all the top hits for 1ei6A were located in the vicinity of the actual Zn^{2+} site even though only a small portion of the structure was correctly modeled. The MetSite scan of the 1jwqA model produced only a single hit with high confidence corresponding to a His residue involved in zinc binding in the native structure. Figures 3.4c/3.4d illustrate the structural predictions are only of moderate quality, both modeled proteins gave MaxSub scores below 30. Nonetheless, residues involved in metal binding were correctly identified for both predicted protein structures.

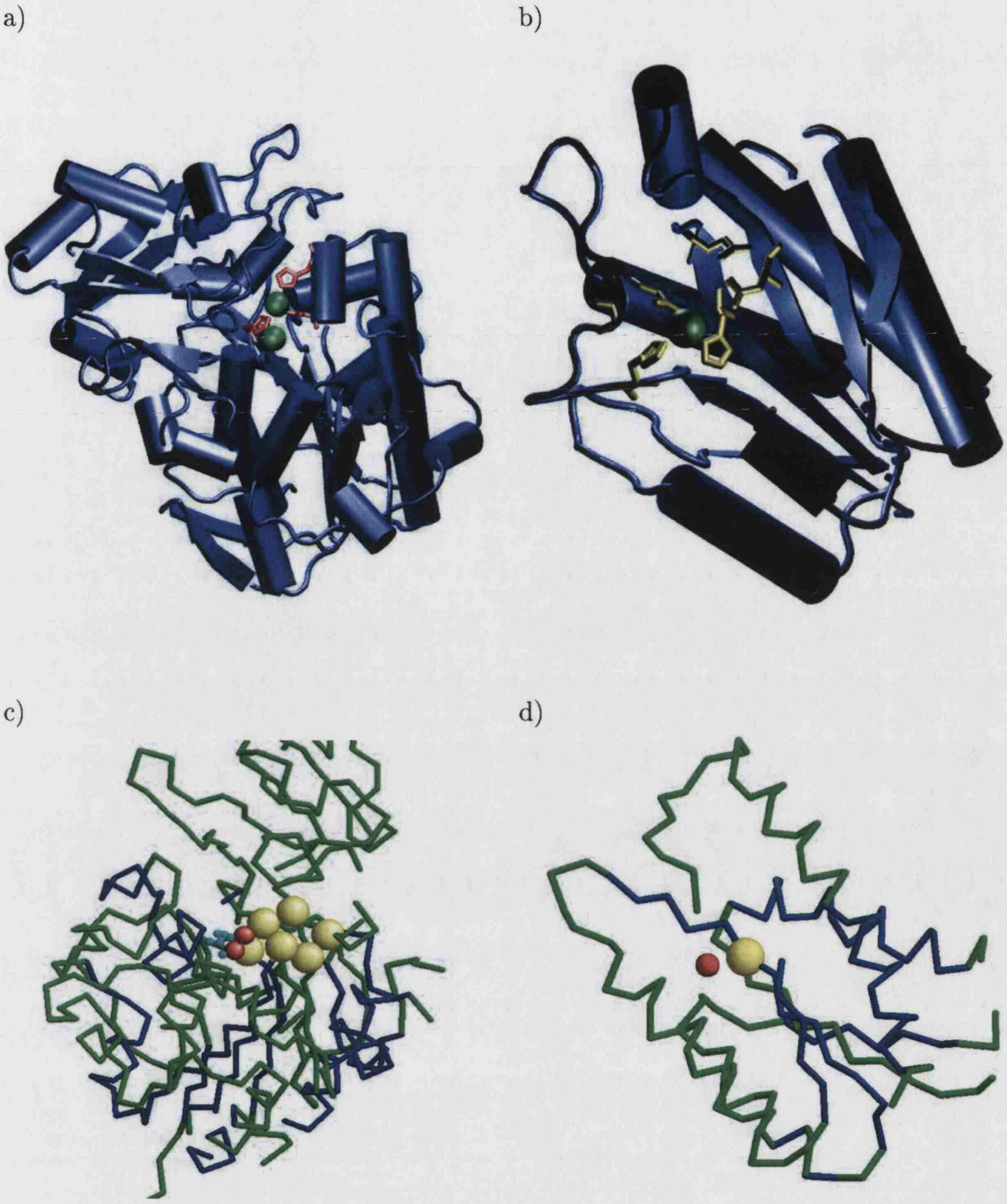


Figure 3.4: MetSite prediction for LiveBench-8 targets. Crystal structures of a) N-acetylmuramoyl-L-alanine Amidase (1ei6A) and b) Phosphonoacetate Hydrolase (1jwqA). The mGenTHREADER predicted structures (blue backbone) shown in c) d) correspond to a) and b) respectively. MetSite predictions are coloured yellow.

3.3.9 Site prediction in a hypothetical protein

Functionally unannotated structures are obviously of particular interest and require accurate methods to correctly identify the location and identity of functional site regions. The LiveBench-8 set contained 42 structures with unknown function annotations. Analysis of the MetSite predictions for these structural genomics targets revealed high scoring hits for the *H. influenzae* hypothetical protein HI0817 (PDB code 1izmA, Galkin et al. (2003)). A BLAST sequence comparison of this target sequence revealed only five hits with an E-value less than 10^{-3} all of which were also hypothetical proteins.

The best mGenTHREADER prediction for this target produced a MaxSub score of only 13 against the NMR structure of Apolipophorin-III (PDB 1eq1A) (Wang et al., 2002). This lipid binding protein is classified as a five-helix bundle belonging to the Apolipophorin-III super family according to SCOP. However, the sequence identity between HI0817 and Apolipophorin-III is only 9.8% and the mGenTHREADER E-value for the structural alignment is 0.068 suggesting that these structures are unlikely to be functionally related.

The Fe^{3+} classifier produced a strong hit against HI0817 with the top five ranking predictions producing network output scores >0.7. In order to visualize the predictions the network outputs were mapped onto the protein structure (Materials and Methods). Figure 3.5 demonstrates all these hits to be clustered around the

same site region corresponding to residues Gly26, Gln105, Asp101, Asn104, Glu23. The overall log likelihood ratio of this site region was calculated to be 13.67 strongly indicating a putative Fe^{3+} binding site.

A) B)

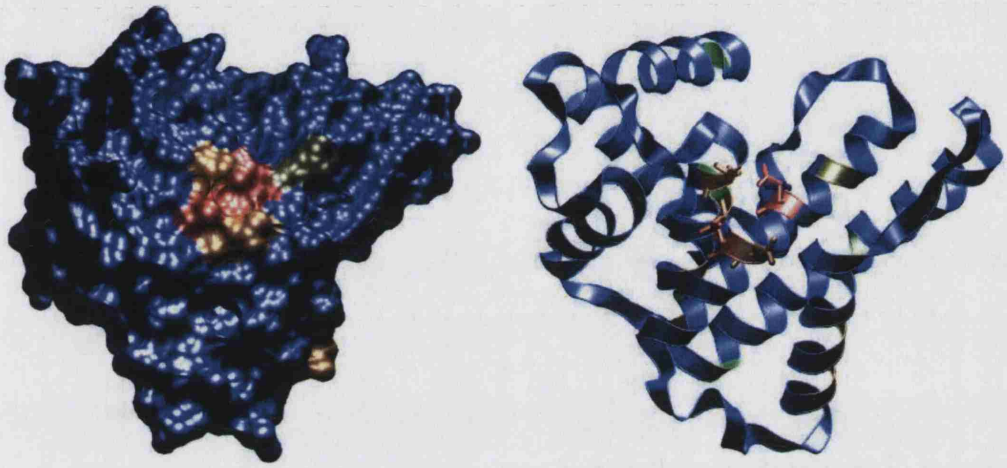
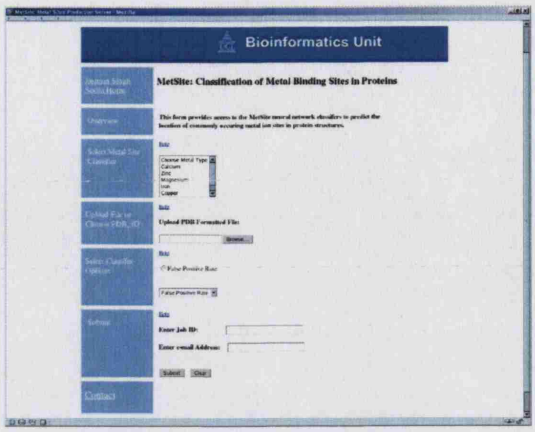


Figure 3.5: MetSite predictions for the crystal structure of structural genomics target HI0817 from *H. influenzae* (PDB code 1izmA). A) Surface representation and B) ribbon representation. The Fe^{3+} neural network scores have been mapped to the temperature factor column in the PDB file, high scoring predictions indicating a putative metal binding site are colored red.

3.3.10 Web Based Predictions

The MetSite classification scheme was developed into a publicly available web-based server. The Java Servlet allows users to upload protein structures in PDB format, which may be modelled proteins, and are asked to provide a set of parameters. Users can specify various acceptable false positive rates according to the level of stringency required. The results are returned as an attachment by e-mail, a PDB format file is returned where the network predictions, above the specified threhold, are mapped to the temperature factor column. Users can view the results by simply coloring by temperature factor in their preferred molecular graphics viewing package. A screen shot of the server and example of the returned structure file is illustrated in Figure 3.6.

A)



B)

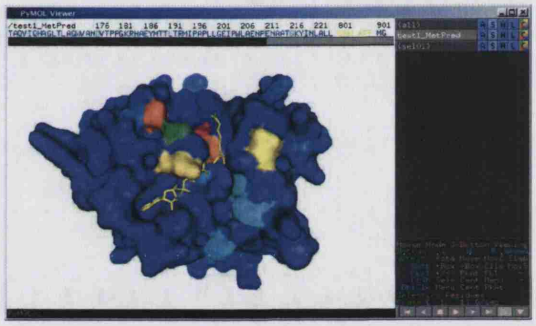


Figure 3.6: Web based access to MetSite classifers. A) users can upload a PDB formatted file selecting desired stringency and B) an example of the results (returned by email attachment as a PDB file) viewed using the Pymol (DeLano (2002)) molecular graphics program and coloured by temperature factor. The example in B) shows the magnesium classifer predictions for a P-loop hydrolase (1a820).

3.4 Discussion

The focus of this chapter has been the development of the MetSite method: a set of artificial neural network classifiers that have been optimised to locate metal binding regions in protein structures. Rigorous benchmarking has shown effective generalisation for identifying metal interacting residues in crystal structures as well accurately predicting metal sites in modelled protein structures from LiveBench.

The results demonstrate that position specific scoring matrices for metal binding residues as well as residues forming second coordination shell interactions, were sufficient for discriminating metal ion sites from non-sites. Secondary structure, solvent accessibility and distance matrices of site residues improve classification performance. The improved classification results obtained by extracting profile scores from residues forming the 3D site, as oppose to residues local in primary sequence, illustrates the importance of implicit spatial encoding. Importantly, the method is restricted to only the approximate location of site residues allowing sites to be identified from homology models where only the approximate backbone conformation is available.

An important aspect of the benchmarking is to ensure no two members of a SCOP superfamily, are present in both training and testing sets thereby mimicking site detection in proteins belonging to new superfamilies. Overall accuracy of site/non-site prediction is high (94.2%) at a low false positive rate of 5%, the method is also

reasonably sensitive at this threshold correctly retrieving 60% of all sites.

Analysis of the most prominent SCOP superfamilies, for each of the metal types investigated, reveal that the method is much more sensitive and selective at identifying metal binding site residues in these structures. On average correct predictions were observed for over 84% of metal sites in these structures with a selectivity of 39%, significantly better performance than over the complete dataset. This is a clear indication that many of the metal sites present in the dataset are likely to be the result of artefacts of crystallisation and as such do not exhibit the sequence/structure characteristics which may indicate a biological significance. Sites may also be occupied by different ion types, for example the calcium binding protein 2pal (Calmodulin) is crystallised with a manganese ion occupying the actual calcium site.

Detection and identification is further complicated by the fact that many active site regions contain more than a single type of metal such as those found in superoxide dismutases. However, a robust approach based on likelihood estimates enables site prediction to be assigned a confidence thereby allowing distinction between different site types.

Validation of MetSite on targets from LiveBench-8 demonstrates the generalisation abilities of the classifiers. These newly released structures do not share any significant sequence similarities to proteins in the PDB, nonetheless for over 71% of these difficult targets, MetSite correctly identified the true metal site in the top

ranking predictions. Accurate predictions were also observed for the active site Mn^{2+} of the RNase domain of the yeast POP2 protein (1uocA), even though the site was not crystallized with the metal present. Manual inspection of the literature revealed that the high scoring hits for this target were indeed biologically important metal sites. Closer inspection of two zinc containing targets (1ei6A and 1jwqA), for which no obvious metal site was detected using the sequence motif searching methods in InterPro, revealed all chelating residues were correctly identified with no false positives.

The MetSite results for a structural genomics targets within the LiveBench set revealed a high scoring hit for the hypothetical protein HI0817 from *H. influenzae* indicating a putative Fe^{3+} binding site. Sequence comparison as well as an InterPro scan of this target sequence did not reveal any obvious functional sites.

Metal site detection was also shown to be possible in the current study for mGenTHREADER predicted structures of the LiveBench set where only approximate backbone position is available. Metal site residues were correctly identified in 8/15 targets for model structures with MaxSub score >0 using only relative backbone coordinates.

The advantages of the machine learning approach presented here are two-fold. Firstly we do not rely upon the specific placement of side chain atoms that are the most prominent ligating donors in the metal sites investigated. The inputs

are restricted for each residue to the position of the $C\beta/C\alpha$ atoms and therefore are more robust to errors in poor quality structures as well as homology models. The second advantage is the speed of the classification; a single protein can be processed within 0.5 sec (using a 1.5Ghz AMD running Linux) given the PSI-BLAST PSSM, enabling genome-wide classification. Furthermore, a robust approach based on likelihood enables site prediction to be assigned a confidence. The method has been implemented into a web-server allowing the research community to access the metal site classifiers.

Given the performance of site detection in modelled proteins it is likely the approach will be useful to complement fold recognition methods. The correct spatial clustering of functionally important residues could be used as a measure of structure prediction quality, this is investigated in greater detail in the last research chapter.

In the next chapter the methods developed in the MetSite approach are extended to determine if functional site detection and classification is accurate for larger site regions. An important challenge is also to utilise the growing number of highly confident model predictions, contained within the Genomic Threading Database (McGuffin et al., 2004a,b) for genome-wide functional analyses.

Chapter 4

Classifying DNA Binding Interfaces and DNA Binding Function

4.1 Introduction

The intricate control of gene expression, as well as the repair and maintenance of genetic material, is achieved through the action of proteins interacting with DNA. It is not therefore surprising that an estimated 6-8% of eukaryotic and 3% of prokaryotic genomes encode DNA interacting proteins (Nadassy et al., 1999; Jones et al., 2001). Transcription factor proteins alone have been estimated to comprise 8% of the human genome (Jones and Thornton, 2004). As a result DNA interacting proteins have been implicated in a number of disease processes making them important targets for pharmaceuticals. Consequently, the analysis of protein-DNA interactions, both experimental and computational, is of great interest and is vital to enhance the understanding of these important classes of proteins.

The aim of this chapter is the extension of the classification system developed for metal sites, in Chapter 3, to the prediction of DNA binding function. The approach is benchmarked for crystal structures and is also assessed for genome-wide structural models.

The requirement for accurate tools to characterise properties of protein-DNA interactions is reflected in the literature: an overview of methods and studies in this area follows.

4.1.1 Mechanisms of DNA Interactions

Generally, computational methods must consider two related but varying problems. In the context of genome-wide annotation, for example, putative DNA binding structures must be distinguished from all other structures. Given a DNA binding prediction, one may require methods to provide detailed predictions of residues forming key interactions with the DNA molecule, perhaps to direct mutagenesis studies to alter the specificity of binding. Understanding the mechanisms by which these interactions occur is therefore of primary importance.

Halford and Marko (2004) recently reviewed the most commonly proposed mechanisms by which selective and efficient interactions with DNA occur. They highlight that diffusion alone is unable to account for impressive ability of DNA binding proteins to identify short stretches of DNA, perhaps only 20 base pairs in length, from the thousands of bases present. Instead of specifically targeting a single DNA site for binding, an alternative suggestion is that the protein may rapidly sample different parts of the DNA molecule by one of three ways: sliding, hopping or intersegmental transfer. The authors argue that such non-specific sampling allows greater coverage of the DNA molecule, significantly reducing binding site search time. An important point regarding intersegmental transfer is that, although the skip size can be up to 400bp, it requires proteins capable of binding DNA at two distinct sites in the protein, such as lac repressor or Sfi1 endonuclease. Such a sampling method would

therefore be prohibitive for most DNA binding proteins where interaction occurs within a deep cleft or ridge within the protein structure.

4.1.2 Binding Specificity

A fundamental property of a large sub-class of DNA binding proteins is the specificity of their interactions. The selective and sensitive nature by which DNA interacting proteins function can have an important influence on many cellular processes and is a vital property of DNA binding proteins. For example, the repair of damage to DNA caused by ultra violet radiation or free radicals, mediated by specific protein-DNA interactions, is immensely important in maintaining genetic integrity and therefore cellular processes. Other important DNA interactions include transcription factor binding, DNA replication and DNA packing, all of which require varying degrees of specificities.

A study of residue conservation in a set of DNA-binding protein families was presented by Luscombe and Thornton (2002). This analysis highlighted that although the overall conservation of DNA interacting residues is indeed higher as compared to non-interacting surface regions the underlying trend is more complex. Residues which form contacts with DNA bases show more variation in conservation between protein families due to varying specificities for different DNA sequence motifs. Proteins may be highly specific for a particular target sequence and therefore the DNA

contacting residues are very well conserved. Conversely, DNA binding proteins capable of binding varying DNA sequences show less conservation of residues interacting with bases.

4.1.3 Structural Analysis

A taxonomy of DNA binding domains was first presented by Harrison (1991) describing the different categories of DNA-binding regions. More recently Luscombe et al. (2000), provided an updated and comprehensive account of the structural features and categories of protein-DNA complexes. The study was based on 240 protein-DNA complexes which were manually classified into eight distinct groups based on structural and functional criteria: Helix-Turn-Helix (HTH), Winged HTH, zinc coordinating, zinc type, β-sheet, β-hairpin, enzyme and other groups. Additionally, the structural alignment program SSAP (Orengo and Taylor, 1996) was used to cluster the complexes into 54 structural superfamilies.

Jones et al. (1999) subsequently presented a detailed analysis of protein-DNA interaction from a structural perspective . Different features of 26 non-homologous protein-DNA complexes were analysed including assessment of both chemical and physical properties of the interface region such as polarity, geometry and packing. The study identified three common modes by which proteins achieve binding to DNA, which the authors termed single-headed, double headed and enveloping.

They highlight that certain common characteristics are required such as an increased tendency for polar residues at the DNA-interface region (contrary to protein-protein interfaces). Interestingly, there is also a higher propensity of water mediated hydrogen bonding. One of the important observations made in this study was the greater prevalence of DNA distortion in complexes where the protein is positioned within the minor groove.

Of the various DNA binding motifs, the Helix-turn-Helix (HTH) is particularly associated with DNA interactions in a variety of organisms. Estimates suggest approximately 1/3 of known DNA binding proteins utilise this motif to interact with the DNA molecule (Jones and Thornton, 2004). Examples include repressor and *Cro* proteins of λ phage, homeodomain and TFIIB in eukaryotes. The Luscombe survey (Luscombe et al., 2000) highlighted that DNA interactions for these proteins generally occur between the major groove bases and the recognition helix of the HTH motif. The second motif helix and linker region provide supporting contacts with the DNA backbone. Although much of the specificity of binding is determined by DNA contacts with the recognition helix the supporting helix may aid specificity. Furthermore, the lack of sequence homology within HTH structural families enables recognition of distinct stretches of DNA. Interestingly, differences are observed between HTH binding proteins of eukaryotic and prokaryotic organisms. For example, the λ Cro transcription factor aligns in the major groove with the helix axis parallel

to base pair edges. In contrast the eukaryotic homeodomain aligns parallel to DNA backbone.

Visualisation tools often allow intricate details of protein interactions to be uncovered. A useful tool, *NUCPLOT*, was developed by Luscombe et al. (1997) to visualise protein-DNA interactions. The program, similar to the popular Ligplot (Wallace et al., 1995), provides a schematic representation of specific atomic contacts including bridging water interactions.

4.1.4 Conformational Changes

An important consideration in any analysis of protein structure and function is the dynamic nature of proteins. Crystal structures represent a time average insight into the conformation of a protein. However, the interaction of biomolecules can result in significant conformational changes (Nadassy et al., 2001). Such changes can even include transition from order to disorder or vice versa.

Jones and Ward (2003) developed a Support Vector Machine (SVM) classifier (DISOPRED) to predict long regions of disorder from protein sequence. Interestingly in a subsequent study (Ward et al., 2004) nuclear proteins, particularly nucleic acid interacting proteins, showed a statistically significant tendency to exhibit disorder regions. This obviously has implications for experimental structure determination, such cases are likely to prove challenging or even impossible to obtain highly

resolved structures through current crystallography techniques.

4.1.5 Predicting DNA Interface regions

Studies have shown that HTH containing DNA binding proteins can be accurately predicted by structural alignment of the query motif to known DNA binding HTH's (Jones et al., 2003a; McLaughlin and Berman, 2003). A root mean squared deviation (RMSD) threshold, of the Cα atoms, of 1.6Å provided a highly selective and sensitive classification scheme. Although this approach is very promising it is limited to structures that contain the HTH motif and has not yet found to be effective for other structural families of DNA binders.

Electrostatic potentials formed the basis of several predictive methods from the Thornton group. In the first study (Jones et al., 2003b) a surface patch based method, originally developed for predicting protein-protein interfaces (Jones and Thornton, 1997), was applied to identify protein-DNA interfaces. Surface residues forming DNA interfaces were labelled by calculating changes in the relative solvent accessibility between complexed and uncomplexed structures. The study concluded that electrostatic potentials were sufficient at discriminating the DNA interface from non-DNA interacting regions in proteins known to bind DNA. Although residue propensity also provided some degree of separation residue conservation and hydrophobicity did not aid the detection of DNA binding sites.

More recently Shanahan et al. (2004) presented a approach which identifies DNA binding regions by combining the template superposition method with electrostatic potentials. The method was demonstrated to be effective for proteins containing the helix-loop-helix, helix-hairpin-helix as well as helix-turn-helix motifs.

4.1.6 Discriminating DNA and Non-DNA Binding

Stawiski et al. (2003) presented a motif independent neural network classification scheme to identify DNA binding sites. Several features were used to train the classifier, including, a hydrogen bonding potential, amino acid composition, residue conservation, surface shape and secondary structure content. Interestingly, the benchmarking was performed on a diverse set of DNA binding proteins against a set of non-DNA binding proteins which exhibit large positive electrostatic surface patches. A useful feature of this approach was highlighted by training the neural network by holding back proteins of the different DNA binding motifs in turn, thereby simulating the prediction for novel DNA binding motifs. Predictions were also presented for a set of known DNA binding proteins crystallised in the unbound form illustrating the effectiveness of structural based approaches for DNA binding detection.

4.1.7 Chapter Overview

The aim of the work presented in this chapter is two-fold. Firstly, given a set of protein-DNA complexes we wish to assess the applicability of the site classification method developed in Chapter 3 to identify residues forming interactions with the DNA molecule. Secondly, given an unknown protein we wish to predict the likelihood of DNA binding.

A fundamental difference of DNA binding sites as compared to metal binding regions is the size and geometry of the interface. DNA binding often involves contacts over large surfaces of the protein and is likely to be present on more exposed areas. We therefore implement a simple clustering scheme to enhance site prediction selectivity by reducing single false positive hits. Results from motif independent testing and for uncomplexed DNA binding proteins suggest DNA binding detection should be possible for novel DNA binding motifs.

The second part of the chapter focuses on the application of DNA binding detection for fold-recognition models generated in the Genomic Threading Database (GTD) (McGuffin et al., 2004a). Gene Ontology (GO) (Harris et al., 2004) annotations are used as a gold standard for the *S. cerevisiae* genome to determine the effectiveness of DNA binding predictions for modelled proteins on a genomic scale. The final part of the chapter focuses on a set of functionally uncharacterised proteins in *S. cerevisiae*. The results for several interesting hits are discussed in more detail.

4.2 Methods

4.2.1 Datasets

The non-homologous dataset of protein-DNA complexes described by Jones et al. (2003b) was used for training and validation. The dataset represents a structurally non-redundant set of 56 protein-DNA complexes.

The dataset presented by Stawiski et al. (2003) was used for the DNA/non-DNA binding discrimination study. This set contains 250 non-DNA binding proteins and 54 DNA binding proteins. The set of non-DNA binding proteins contains proteins which exhibit large positive electrostatic patches on the surface similar to those expected to form at DNA binding interfaces.

4.2.2 Defining Interface Residues

Naccess (Hubbard and Jones, unpublished) was used to calculate solvent accessibility of both the complexed and uncomplexed PDB coordinates. Residues were labelled as DNA interacting if the relative solvent accessible area decreased by $>1Å^2$ on binding. The Naccess method implements the algorithm described by Lee and Richards (1971) where a probe is 'rolled' over the protein surface. The default probe size of 1.4Å (radius of a water molecule) was used.

4.2.3 Neural Network Training

The Matlab neural network toolbox was used to train the neural networks to classify the DNA site data. A simple feed forward architecture was chosen consisting of three layers: an input layer, a hidden layer and an output layer with a single node. In order to acheive effective generalisation it is important to limit the number of weights in the neural network in relation to the number of examples used in training. The number of hidden nodes was set using the simple rule: $\sqrt{2a_i}$, where a_i is the number of input nodes for network i. All networks described here were trained using the resilient backpropagation algorithm Riedmiller and Braun (1993) with early stopping using the transfer functions described previously in section 3.2.4.

4.2.4 Assessment Metrics

The cross-validation results for both the residue based and site based predictions are assessed in a similar manner as previously described in Chapter 3. The Q_2 accuracy is defined as the total number of true positive and true negatives over all patterns. Sensitivity (TPR), false positive rate (FPR) and selectivity are defined as follows:

$$\text{TPR} = \frac{(TP)}{(TP + FN)} \tag{4.1}$$

$$\text{FPR} = \frac{(FP)}{(FP+TN)} \tag{4.2}$$

$$\text{Site Selectivity} = \frac{(TP)}{(TP+FP)} \tag{4.3}$$

where T = True, F = False, P = Positive, N = Negative and R = Rate.

4.2.5 Site and Patch Prediction Clustering

An important distinction between metal ion sites and larger functional site regions such as DNA binding interfaces is the shape and size of sites. The DNA interface has been shown to span, on average, 1600Å^2 in a large scale analysis of protein-DNA interactions (Nadassy et al., 1999). A post-processing clustering step is therefore introduced for DNA binding predictions which takes into account a larger region of the predicted area as DNA binding. Such clustering is likely to provide a better representation of the predicted interface whilst eliminating spurious hits.

The electrostatic based method described by Shanahan et al. (2004) presented a patch based approach to define DNA interfacing regions. The patches were defined by taking every residue on the protein surface as a starting point and then extended the patch by identifying the nearest neighbour with a positive electrostatic score. Patch sizes were limited to 10 surface residues, with the definition for a correct

prediction requiring 7/10 residues in the top ranking patch to be true DNA interface residues.

In the current approach the neural network output for each surface residue is used to predict protein-DNA interface patches. Patches are defined and extended in a similar way as described in the Jones et al. (2003b) study. However, instead of moving the patch in the direction of the positive electrostatic score, in the current study we move toward residues with neural network scores greater than a network threshold. Patches are again limited to 10 residues and patch scores calculated as the sum of neural network outputs for a given patch.

4.2.6 Genome-Wide Binding

Protein structures for the genome-wide analysis of DNA binding function were derived from the Genomic Threading Database (GTD) McGuffin et al. (2004a). The GTD contains structural assignments by GenTHREADER from a number of organisms. Confidence values are assigned to each of the predictions. The DNA classification scheme was incorporated into the distributed system used for the GTD (collaborative work with McGuffin, L.J) in order to obtain DNA biding predictions for the model proteins.

4.2.7 Benchmarking Genome Function Assignment

In order to determine the effectiveness of DNA binding predictions on a genome-scale, the cross-validated neural networks were used to classify structural predictions derived from the *S. cerevisiae* genome. The Gene Ontology was used to label the proteins in this dataset. All sequences annotated with molecular function DNA binding (or a child of this term) were taken to be the positive DNA binding set. All proteins not within the DNA binding set were taken as the control negative dataset provided they did not also belong to molecular function 'unknown' group.

4.2.8 Identifying DNA Binding Motifs

DNA binding function is known to be performed by a number of different structural motifs. A number of approaches in the literature have focused on identifying proteins which use the Helix-Turn-Helix motif to bind DNA. However, such approaches are restricted in their scope, depending on the presence of such motifs. Cross-validation was performed in the Stawiski et al. (2003) study by grouping together DNA binding proteins by structural motif. Training was performed by leaving out all members of a given structural motif group in turn to assess the ability to detect DNA binding for novel motifs. This cross-validation experiment was repeated, using the neural networks developed in the current study.

4.3 Results

4.3.1 Prediction of DNA Binding Residues

The classification performance for the prediction of residues involved in DNA binding using the encoding and training scheme developed previously (Chapter 3) was initially assessed. The dataset of 56 protein-DNA complexes were used to perform leave one out jack-knifing validation for each protein in the set. The jack-knifing was repeated using the different features used in the site encoding. The Naccess derived definition for DNA binding residues (Methods) resulted in a total of 2021 residues, out of a total of 11287 to be labelled as DNA interacting.

Overall, incorporating all structural and sequence features provided the best classification performance. From the cross-validation results in Table 4.1, at a fixed false positive rate of 5%, we see the sensitivity of binding residue detection is 26.9% with a selectivity of 60.7% when incorporating all 285 features. This compares to 25.1% and 59.0% when using only PSSM scores of residues local in 3D space. These findings are consistent with previous results from metal binding site detection in that the majority of the discrimination is encoded within the vector of profile scores.

Feature Set	Network Cut-Off	Q2 Accuracy (%)	TPR (%)	Sel (%)	Wilcoxon
All (285)	0.56	80.7	26.9	60.7	0.77
Site PSSM (200)	0.55	79.4	25.1	59.0	0.76
Sequence PSSM (200)	0.54	78.9	24.1	58.4	0.74
SS Only (30)	0.36	74.7	4.1	19.0	0.52

Table 4.1: Overall cross-validation results for classifying DNA binding residues using different feature sub-sets. Neural network cut-off represents the network threshold at a fixed 5% FPR. PSSM = PSI-BLAST position specific scoring matrix, SS = DSSP secondary structure assignment. All features represent PSSM, Solvent Accessibility, Distance Matrix, and SS features combined. Site and Sequence PSSM relate to residues local in 3D-space or local in sequence space respectivley.

4.3.2 Patch Based Predictions

Table 4.2 presents the leave-one out jack-knifed predictions for the dataset of 56 protein-DNA complexes. The results have been ordered by the number of residues in the top ranking patch (as described in the methods) which are DNA interfacing residues. The results also include the residue based classifications: all surface residues which produce a network output above a threshold corresponding to less than 5% false positives are predicted to be DNA interfacing. True positives (TP) and false positives are then determined.

Overall, 35/56 (62.5%) of the proteins in the dataset contain $\geq$7 interface residues

in the top ranking patch of 10 residues. Furthermore, 1023 of the 2021 (50.6% sensitivity) DNA interface residues were correctly identified with a selectivity of 31%. The overall accuracy (Q_2) for locating interface residues was calculated to be 69%.

These results are encouraging and highlight effective DNA binding/interface predictions using the extended MetSite methods. The results are comparable to the findings from the Jones study in which 38/56 (68%) proteins were correctly predicted, using the equivalent definition of a correct prediction (7/10 residues in top patch must be true interface residues).

4.3.3 Discriminating Contact Type

The characteristics of protein-DNA interactions vary depending on the biological context of the interaction. For example, transcription factors which bind DNA at very specific DNA sequence motifs are expected to make more selective contacts with the bases of DNA. Conversely, histone proteins generally form more non-specific contacts with the DNA backbone. Luscombe and Thornton (2002) presented a study of conservation of DNA interfacing residues and found some intriguing results. Generally residues contacting the DNA backbone were more conserved as compared to residues forming contacts with the bases of the DNA. This is most likely due to variations in binding specificity of base contacting residues whilst overall stabilisation

PDB ID Code	Protein Name	Interface Residues	Non-Site Residues	TP	FP	Top Patch	Sens (%)	Selc (%)	RPV
1bp7A	Endonuclease I-CREI	62	44	35	13	10	56.45	46.67	0.41
1eonA	Type II restriction enzyme ECORV	46	111	15	29	10	32.61	20.00	0.13
1eqzA	Histone H2A	33	68	29	30	10	87.88	46.03	0.06
1hcrA	HIN recombinase	23	16	16	3	10	69.57	61.54	0.49
1ignA	RAP1	55	75	33	14	10	60.00	47.83	0.33
1pdnC	PRD paired domain	53	52	36	31	10	67.92	42.86	0.21
1xbrA	Transcription factor T domain	29	94	25	18	10	86.21	53.19	0.14
2cgpA	Catabolic gene activator protein	17	114	11	0	10	64.71	64.71	0.03
6croA	Lambda CRO	30	18	20	10	10	66.67	50.00	0.09
1au7A	PIT-1 POU domain	44	45	24	15	9	54.55	40.68	0.22
1c9bA	Transcription factor IIB	41	86	24	8	9	58.54	48.98	0.18
1crxA	Cre recombinase	65	154	37	18	9	56.92	44.58	0.20
1ecrA	Replication terminator protein (TUS)	67	131	34	23	9	50.75	37.78	0.14
1gdtA	Gamma-delta resolvase	71	67	28	10	9	39.44	34.57	0.23
1ihfA	Integration host factor	63	16	29	12	9	46.03	38.67	0.23
1qnaA	Transcription initiator factor TFIID-1	43	75	31	4	9	72.09	65.96	0.19
1sknP	SKN-1 transcription factor	21	33	15	4	9	71.43	60.00	0.24
2irfJ	Interferon regulatory factor-2	27	50	18	9	9	66.67	50.00	0.19
1azpA	SAC7D	20	21	11	2	8	55.00	50.00	0.38
1emhA	Uracil-DNA glycosylase	18	102	14	29	8	77.78	29.79	0.09
1qumA	Endonuclease IV	30	104	17	11	8	56.67	41.46	0.09
1tauA	DNA polymerase	42	379	25	64	8	59.52	23.58	0.03
1vasA	Endonuclease V	43	52	18	11	8	41.86	33.33	0.17
2bdpA	Nucleotidyl Transferase	63	243	44	36	8	69.84	44.44	0.07
1a36A	DNA topisomerase	68	272	47	73	7	69.12	33.33	0.11
1am9A	Sterol regulatory element binding protein 1A	17	50	12	13	7	70.59	40.00	0.17
1b3tA	EBNA-1 nuclear protein	31	73	26	45	7	83.87	34.21	0.21
1dizA	3-Methladenine DNA glycosylase	51	123	24	18	7	47.06	34.78	0.09
1dp7P	RFX-DBD (Transcription)	26	35	14	16	7	53.85	33.33	0.09
1gd2E	BZIP transcription factor RAPI	14	36	11	5	7	78.57	57.89	0.15
1hwtC	HAP1	18	33	10	22	7	55.56	25.00	0.18
1lmb3	Lambda repressor	16	7	14	9	7	87.50	56.00	0.23
1qpiA	Tetracycline repressor	21	111	8	19	7	38.10	20.00	0.01
1tupB	Tumour supressor P53	26	96	14	64	7	53.85	15.56	0.04
6mhtA	HHAI methyltransferase	40	149	8	8	7	20.00	16.67	0.17
1qpzA	Purine repressor	14	169	12	23	6	85.71	32.43	0.00
2bopA	E2 DNA-binding domain	11	53	9	31	6	81.82	21.43	0.01
1ewnA	AAG DNA repair glycosylase	22	95	17	25	5	77.27	36.17	0.15
2dnjA	Deoxyribonuclease 1	21	113	11	33	5	52.38	20.37	0.08
1a1hA	QGSR zinc finger	35	32	28	26	4	80.00	45.90	0.25
1bdtA	Arc transcription regulator	26	18	0	0	4	0.00	0.00	0.31
1qrvA	Endonulcease V	13	16	5	5	4	38.46	27.78	0.14
3htsB	Heat shock transcription factor	13	43	8	14	4	61.54	29.63	0.06
3pviA	Endonulcease PVUII	45	54	5	7	4	11.11	9.62	0.17
1d02A	Restriction endonuclease MUNI	31	88	15	19	3	48.39	30.00	0.09
1dfmA	Restriction endonuclease BGII	75	64	16	18	3	21.33	17.20	0.26
1fokA	Restriction endonuclease FOKI	66	242	30	59	3	45.45	24.00	0.12
1mjoB	Methionine repressor	31	42	10	10	3	32.26	24.39	0.10
2hmiA	HIV-1 reverse transcriptase	57	303	20	102	3	35.09	12.58	0.03
1a3qA	NF-KAPPA-B	21	166	9	42	2	42.86	14.29	0.04
1dctA	DNA (cytosine-5) methylase	49	104	5	21	1	10.20	7.14	0.04
1dmuA	Restriction endonuclease BGLI	40	142	1	24	1	2.50	1.56	0.05
1a73A	Endonuclease I	55	68	21	31	0	38.18	24.42	0.20
1bg1A	STAT3β	15	315	7	81	0	46.67	7.29	0.01
1mhdA	SMAD MH1 domain	20	51	7	19	0	35.00	17.95	0.11
1zqfA	DNA polymerase β	27	169	10	34	0	37.04	16.39	0.02

Table 4.2: Cross-validation results for classifying DNA binding residues. Top patch represents the number of interface residues in the top ranking patch prediction. The Random Prediction Value (RPV) is the probability of predicting a patch containing at least 7 DNA interfacing residues by chance.

of the interaction is performed through backbone contacts.

The patch based classification results were analysed in order to determine the number correctly predicted backbone (BB) or base-pair (BP) contacting residues. Residues are defined to be base-pair contacting if they are observed to form more contacts with base-pair atoms as compared to DNA backbone atoms. Patch scores are calculated using the algorithm outlined above and interface predictions are taken if a patch score is above the expected 5% false positive threshold. The results indicate that the accuracy of identifying base-pair interfacing residues is higher as compared to backbone interactions: of 318 BP interface residues, 207 are correctly predicted (65.1%) as compared to 910/1657 (54.9%) of BB contacting residues.

4.3.4 Discriminating DNA Binding Function

An important issue for genome annotation is not only the ability to discriminate between interface and non-interface residues of proteins which interact with DNA but also the effective discrimination of non-DNA binding proteins. Methods based solely on identifying electrostatic patches will lead to a high number of false positives (Jones and Thornton, 2004; Stawiski et al., 2003) as many non-DNA interacting proteins also exhibit large positive electrostatic patches on the protein surface.

The residue classification study was extended to determine the effectiveness at discriminating DNA binding proteins from a dataset of non-DNA binders. Stawiski

and coworkers presented a dataset of 250 non-DNA binding proteins and 54 DNA binding proteins (Stawiski et al., 2003). Crucially the selection of the negative set included proteins exhibiting large positive electrostatic patches on the surface. This dataset was used in the current study for benchmarking DNA/non-DNA binding discrimination.

Non-homologous protein chains from the dataset were randomly split into five groups to perform cross-validation. Training was performed using all sequence and structural features (285 inputs). The top ranking patch score, using the cross-validated weights, was taken as the prediction for DNA binding for a given protein.

The distribution of top ranking patch scores for the DNA binding and non-DNA binding datasets is illustrated in Figure 4.1. The results clearly illustrate a distinct separation between the datasets. The mean of the top ranking patch score for DNA binding proteins was calculated to be 6.4(±)1.6 as compared to 4.1(±)1.5 for the non-DNA binding set.

An overall Q_2 accuracy of 87.8% was achieved using a patch score threshold of 5.7. The sensitivity (TPR) at this threshold is also high: 38/54 (70.4%) of the true DNA binding proteins are correctly predicted, whilst only 25/250 (10%) of the non-DNA binders are incorrectly classified.

Figure 4.2 illustrates ROC plots for the classification results. For comparison the classifications using the above neural network with either patch scores or site

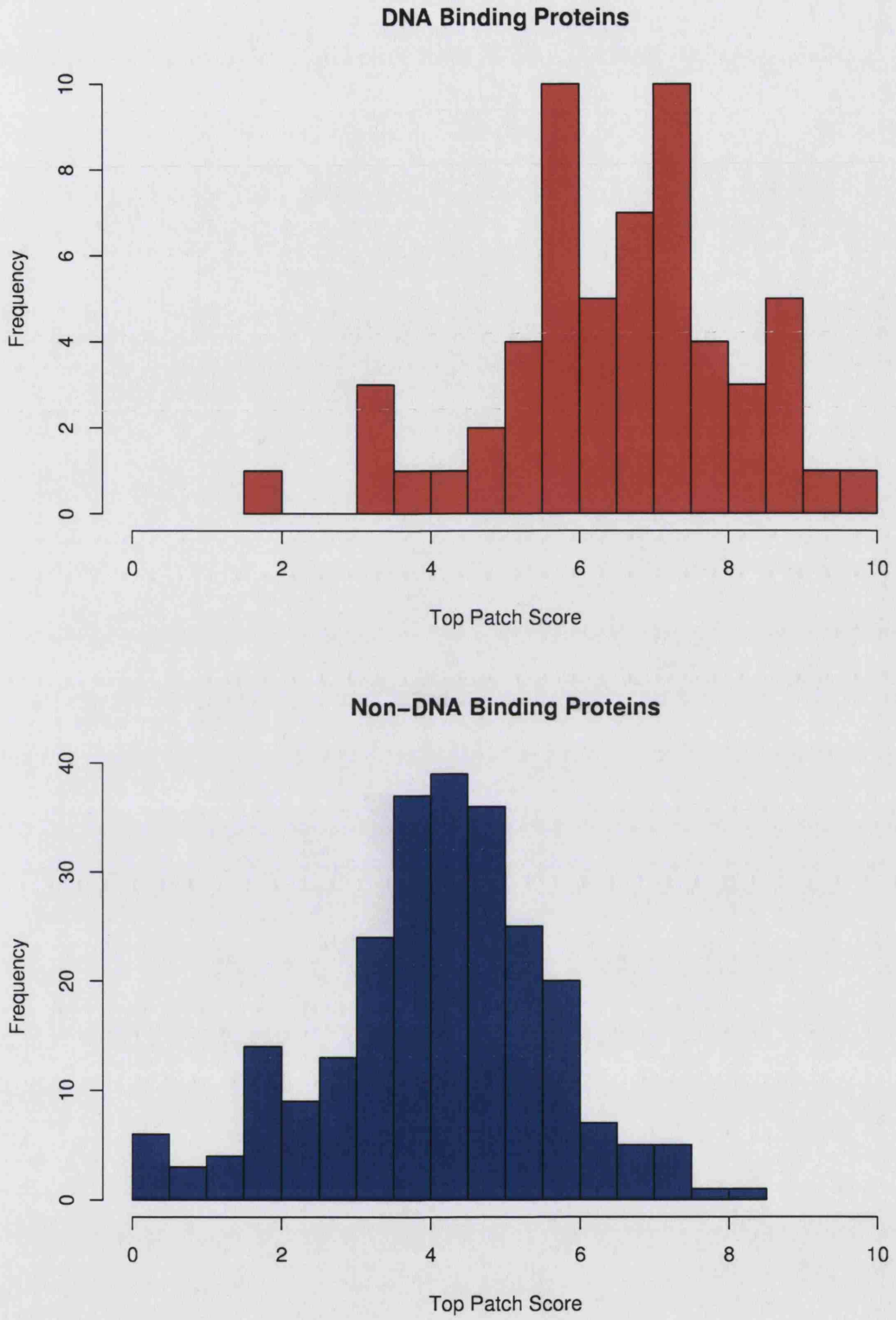


Figure 4.1: Distribution of top ranking DNA Site Predictions: A) DNA binding proteins and B) non-DNA binding proteins.

scores used to classify DNA binding function have been included. In addition a simple neural network classifier was trained using only sequence composition for each protein chain in the dataset. The most accurate classification is observed for the results described above where training is performed using only surface exposed residues and classifying DNA binding function based on top patch score. Scoring the hits using a site based system, where only the closest neighbours of a residue contribute to the site score rather the patch extending algorithm, resulted lower sensitivity (63% at 10% FPR). The simple sequence composition classifier resulted in a sensitivity of 53.7% at this FPR.

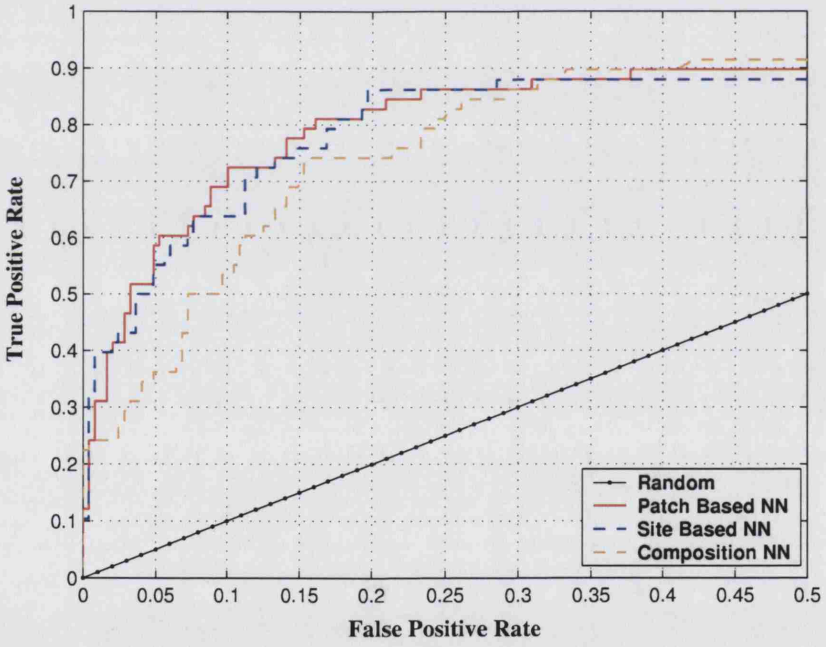


Figure 4.2: ROC curves for DNA/non-DNA binding discrimination using patch or site post-processing or a simple neural network trained on sequence composition.

4.3.5 Structural Motif Based classification

Commonly occurring local DNA binding motifs were characterised by Jones et al. (2001). The study identified eight major groups which share important structural features in the DNA interacting region. In order to determine the performance of the current classification approach for detecting novel binding motifs, cross-validation experiments were performed by grouping the dataset by structural binding group. The neural network is trained by leaving out each structural motif group in turn thereby mimicking classification performance for novel DNA binding motifs. This type of analysis was also presented by Stawiski and co-workers (Stawiski et al., 2003). The results are presented in Table 4.3.

Binding Motif	**Total Chains**	**Patch Score $\geq$5**	**Average Patch Score**
HTH	16	11	5.4±1.1
Enzyme	16	6	5.0±1.1
β-Sheet and β-ribbon	7	1	3.2±1.8
Other α-helix	7	6	6.4±1.0
Zinc-finger	4	4	6.2±0.8
Zipper-type	2	2	7.2±0.8
Other	2	1	4.6±0.8

Table 4.3: Classification results for motif independent training (*as described by Stawiski et al. (2003)*)

The overall motif cross-validation tests show 31/52 structures produced top ranking patch scores $\geq$5. The zinc finger and zipper group only contained 2 and 4 structures respectively, however all structures with these motifs produced patch scores above the threshold of 5. Performance was also good for the HTH group. In total 11 of the 16 structures in this set were correctly predicted to bind DNA. The enzyme group represents a special set of cases: proteins in this set all display enzymatic activity when bound to DNA. Of the 16 enzyme cases, however, only 6 are predicted as DNA binding. The β-Sheet and β-ribbon cases showed the worst success rate: only 1 of the 7 proteins in this set were correctly predicted as DNA binding.

Figure 4.3 provides a set of examples from the motif cross-validation results. The neural network predictions were mapped to the B-factor column of each PDB file (as described in Chapter 3). The structures were prepared for viewing using the Pymol molecular graphics program (DeLano, 2002) and residue colouring preference set to temperature factor option. The results illustrate that high scoring hits generally tend to be grouped in regions close to the DNA interface.

The HHAI methyltransferase protein from the enzyme group is illustrated in Figure 4.3a. The crystal structure of this protein shows a DNA base 'flipped' into the active site of the protein. Residues within the helix facing the DNA molecule, as well as the turn regions interacting with the DNA produced several strong hits.

The classification results for the HTH containing λ-repressor protein is depicted

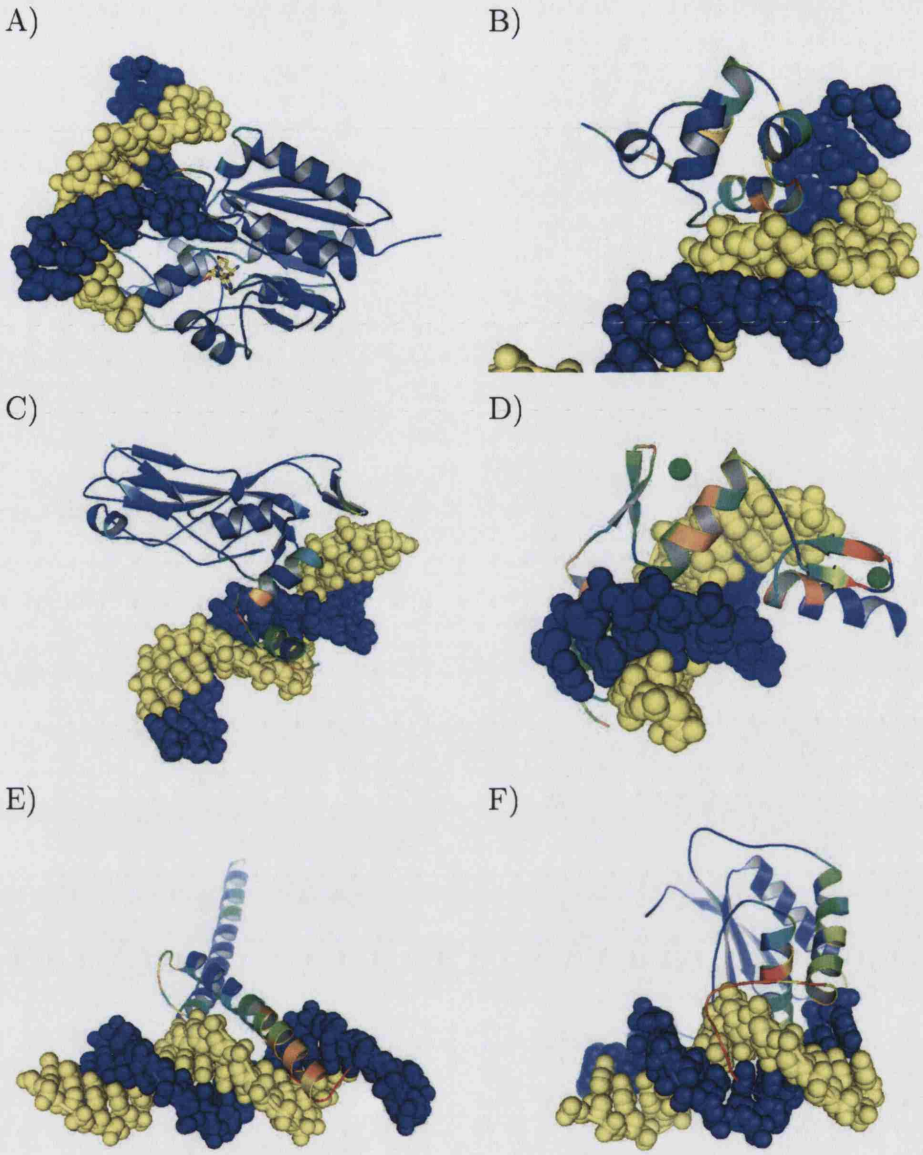


Figure 4.3: Motif Independent Classification of DNA binding regions for proteins from each structural binding group. A) HHAI methyltransferase (3mhtA, enzyme), B) λ-repressor (3croL, HTH), C) Transcription factor (1bxrA, β-type), D)Zinc finger complex (1aayA, zinc-finger), E) Transcription regulation (1am9C, zipper-type), F) Nuclear protein (1b3tA, α-helix other). The residues have been coloured by neural network output which have been used to replace the B factor values of the PDB file.

in Figure 4.3b. The recognition helix located in the major groove of the DNA clearly produces several hits, particularly strong hits are also observed for two residues located in the turn regions flanking the helix. However, the N-terminal helix, which does not seem to interact with the DNA, also produces a number of hits, albeit of a lower network output.

The transcription factor in Figure 4.3c was the only structure producing a high site score for the β-type group. The two N-terminal α-helices with the connecting loop produced the highest scoring hits in this structure. Interestingly, a moderate score is also observed for a β-hairpin facing the DNA.

Figure 4.3d shows a zinc-finger/DNA complex. The regions coloured red and orange correspond to particularly high network outputs (≥ 0.7) and correspond to areas involved with Zn^{2+} biding as well as the α-helix located in the major groove of the DNA molecule.

The results for the transcription regulator protein (zipper type) in Figure 4.3e demonstrates the high scoring hits located on the N-terminal helix lying in the DNA major groove. Protein DNA contacts are prominent for the initial 22 residues of the protein, all of which correspond to high network values. High scores are also observed for residues in the linker region which also make important contacts with the DNA.

An example from the α-helix other group is presented in Figure 4.3f, and depicts

the DNA binding domain of the nuclear protein from the Epstein-Barr virus. Interestingly critical DNA contacts are made by an N-terminal loop region which wraps around the DNA molecule. This loop corresponds to several high scoring hits: 11 residues with an average network output of 0.75 (site score of 7.88). Intriguingly, the conformation of the loop suggests a possible mechanism for the DNA interaction, perhaps involving a shift in the protein structure to achieve more selective binding.

4.3.6 Predicting DNA Binding Proteins in their Unbound State

The interaction of proteins with DNA can result in significant conformational changes and even disorder to order transitions (Nadassy et al., 1999; Ward et al., 2004). Stawiski et al. (2003) presented a dataset of unbound proteins which are known to bind DNA. The cross-validated neural network weights were used to determine the accuracy of DNA binding prediction in these unbound structures. The unbound set contained proteins which were homologous to proteins in the training set. For such cases neural network weights derived from training performed by excluding the homologous protein were used for classification. The results are shown in Table 4.4.

Patch scores greater than 5 were observed for 7/11 of the unbound structures. The highest scoring prediction was observed for the Homoeodomian single chain protein (1enh). Of the novel DNA binding motif proteins, the human Sp100B SAND

(1h5p) domain produced the highest scoring DNA site prediction.

PDB Code	Protein Description	Protein Classification	Top Patch Score	Resolution (Å)	Homolog in Training Set
1enh0	Engrialed homeodomain	Homeodomain	11.8	2.1	None
1hcp0	Estrogen Receptor	Zinc-finger	7.3	NMR	2nllA
1etc0	ETS domain	Winged helix	6.73	NMR	1bc8C
2cro0	434 cro repressor	HTH	6.59	2.3	3croL
1h5p0	Sp100B SAND domain	Novel	5.98	NMR	None
1tbpA	TATA-binding protein	β-ribbon	5.93	2.6	1volB
1dm9B	Heat shock protein 15	Novel RNA binding	5.19	2	None
1bm80	Mbp1 protein	Winged helix, novel	4.72	1.7	None
1c8z0	Tubby protein	Novel	4.64	1.9	None
1bix0	DNA repair endonuclease	Enzyme	3.49	2.2	None
2hts0	Heat shock protein	HTH variant	2.74	2.2	3htsB

Table 4.4: Classification results for set of protein chains known to bind DNA but uncomplexed in the crystal structure. Dataset from the Stawiski et al. (2003) study

Analysis of InterPro revealed that the SAND domain is a conserved region found in a number of nuclear proteins, which have been shown to be involved in chromatin-dependent transcriptional control (Mulder et al., 2002; Gibson et al., 1998). This domain has also been implicated in various human diseases. The DNA binding function of this domain was determined, through mutagenesis studies, to be performed by a conserved KDWK motif (Bottomley et al., 2001). This motif is located at position 653-656 of the Sp100b SAND sequence. The top ranking DNA site prediction in the unbound structure was centred on residue Ser 658, which is within 6.5Å of the actual DNA binding residues of the KWDK motif. These findings are encouraging as not only does the domain bind DNA via a novel region, it is uncomplexed and

also a NMR structure.

An important consideration with the above analyses is that the dataset is very limited. However, a more complete benchmarking study of DNA binding proteins in unbound form is not currently possible due to limited number of examples currently in the PDB (Jones and Thornton, 2004). Future additions, especially from structural genomics projects, are likely to address this issue.

4.3.7 Gene Ontology Assignments

The DNA classification results for the *S. cerevisiae* genome were assessed against GO annotations (Methods). The results indicate the majority of DNA interactions in *S. cerevisiae* are described as DNA polymerases (41%) followed by transcription factor activity (19%). RNA polymerases make up the third largest group followed by DNA helicase and chromatin binding. The two smallest groups in the set were labelled as DNA dependent ATPases and ssDNA binding. Figure 4.4 summarises the proportion of protein sequences for the different GO annotations.

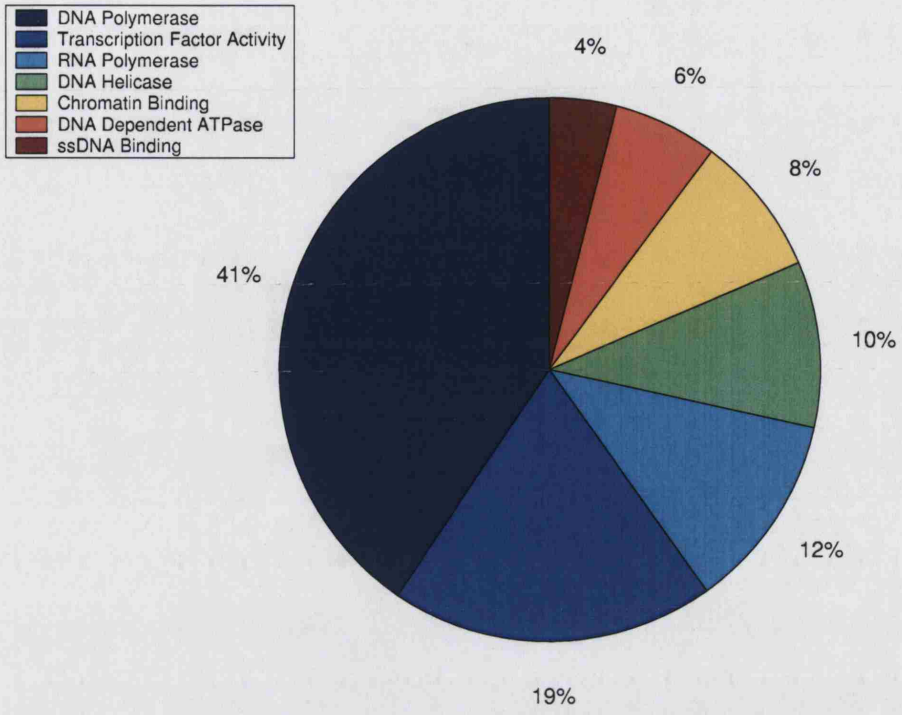


Figure 4.4: Proportions of sequences from the *S. cerevisiae* genome annotated with descendents of the term 'DNA binding' within the molecular function ontology.

4.3.8 Assigning Confidence

An important consideration for genome wide functional assignments is an effective means to assign putative hits a confidence value to allow the significance to be determined. In order to determine confidence values, the single best site score from the top 10 ranking protein models, generated by GenTHREADER, was extracted for the each DNA binding and non-DNA binding protein. This score was taken to

be the prediction for a given protein to interact with DNA. The predictions were assessed against the GO labels to calculate the reliability measures at varying site score thresholds. Figure 4.5 illustrates different measures of reliability at varying site score values. The results indicate that, at a site score of over 11, one would expect only a 1% likelihood of a false hit.

4.3.9 Genome-wide DNA Binding Prediction

The GO annotations were used to perform classifications of DNA binding on structural models for the *S. cerevisiae* genome. Of the 5864 unique protein sequences in the *S. cerevisiae* genome, 844 homologies were identified with proteins in the DNA binding classifier training set. In order to ensure the classification results were rigorously benchmarked, cross-validated weights were used for sequences exhibiting BLAST E-values $\leq$0.1 to any protein used to train the DNA binding classifier. The number of hits, at different false positive thresholds, for the various GO classes is presented in Table 4.5.

Setting the site score threshold to a very stringent expected false positive rate of 1% resulted in 44/198 (22.2%) of protein models labelled as DNA binding to be retrieved correctly with only 16 false positives (73.3% selectivity). Sensitivity was improved by relaxing the site score threshold to a 5% false positive rate, allowing 68/198 (34.4%) to be retrieved with 104 false positives (39.5% selectivity).

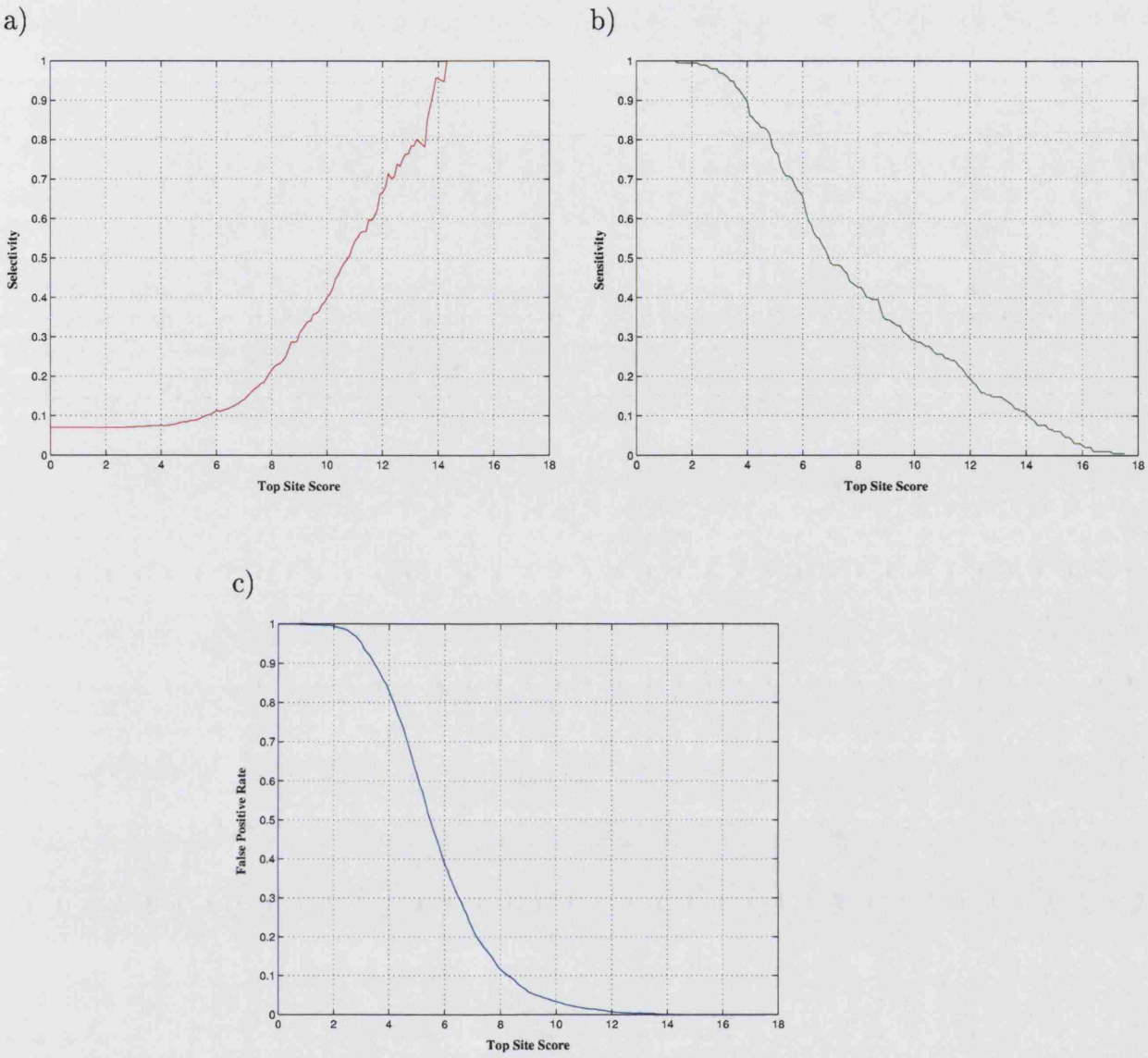


Figure 4.5: Site score reliability measures for DNA classification of known DNA and non-DNA binding proteins as annotated by the Gene Ontology molecular function for the *S. cerevisiae* genome. Plots show the a) selectivity ($TP/(TP+FP)$), b) Sensitivity ($TP/(TP+FN)$) and c) false positive rate ($FP/(FP+TN)$) against DNA site score.

Gene Ontology Molecular Function Term	Total Number	1% FPR	5% FPR	10% FPR
DNA Binding	198	44	68	83
Polymerase	74	2	10	14
Transcription Factor Activity	53	23	29	31
DNA directed RNA Polymerase	32	0	4	5
DNA Helicase	27	0	2	6
Chromatin Binding	23	2	4	10
DNA Dependent ATPase activity	17	0	2	8
Single Stranded DNA binding	11	0	0	1
Non-DNA	2177	16	104	216

Table 4.5: Breakdown of the Gene Ontology annotations at varying false positive rates (FPR) for *S. cerevisiae* containing the keyword DNA in the molecular function category. The total number gives the total number of proteins annotated to a particular GO term.

On the whole, analysis of the individual classes of GO terms showed less accurate classifications. The transcription factor activity class, however, was an exception: 29/53 of these proteins were accurately predicted to bind DNA from the model structures.

4.3.10 Annotating Unknown Functions

All sequences annotated with GO molecular function term 'unknown function' were extracted from the dataset giving 2222 protein sequences. These were screened

against the DNA/non-DNA classification training set. Interestingly 185 homologies to proteins in the training set were detected with BLAST evalues ≤ 0.1. Of these 32 were to DNA binding proteins and 153 to proteins from the non-DNA binders. This highlights a general issue with annotations in GO: in many instances it may be that the individual performing the annotation has not performed simple homology searches when assigning the molecular function. Thus functional annotations to close homologs may be missed. Alternatively, a more stringent confidence level may be set. For each protein, models were generated for the top ten GenTHREADER structural alignments. DNA binding site predictions were then performed. The top ranking site score, across all ten models, was taken to be the prediction for a given sequence. In order prevent over optimistic site predictions, the homologous cases were extracted from the dataset and classified using cross-validated weights. Figure 4.6 shows the histogram of top site score for protein generated from the unknown set. Note this plot has been filtered to remove the simple homology cases.

At a very stringent level of confidence, setting the score threshold to 11.6 which corresponds to an expected false positive rate of 1%, we observe two hits which are predicted to bind DNA. Relaxing the threshold to a 5% FPR results in more hits, 30 in total which may be of some significance. Analysis of the homologous cases reveals a much higher hit rate at equivalent thresholds. Of the 32 DNA binding homologs, 17 of the GO annotated unknown function proteins are predicted to bind

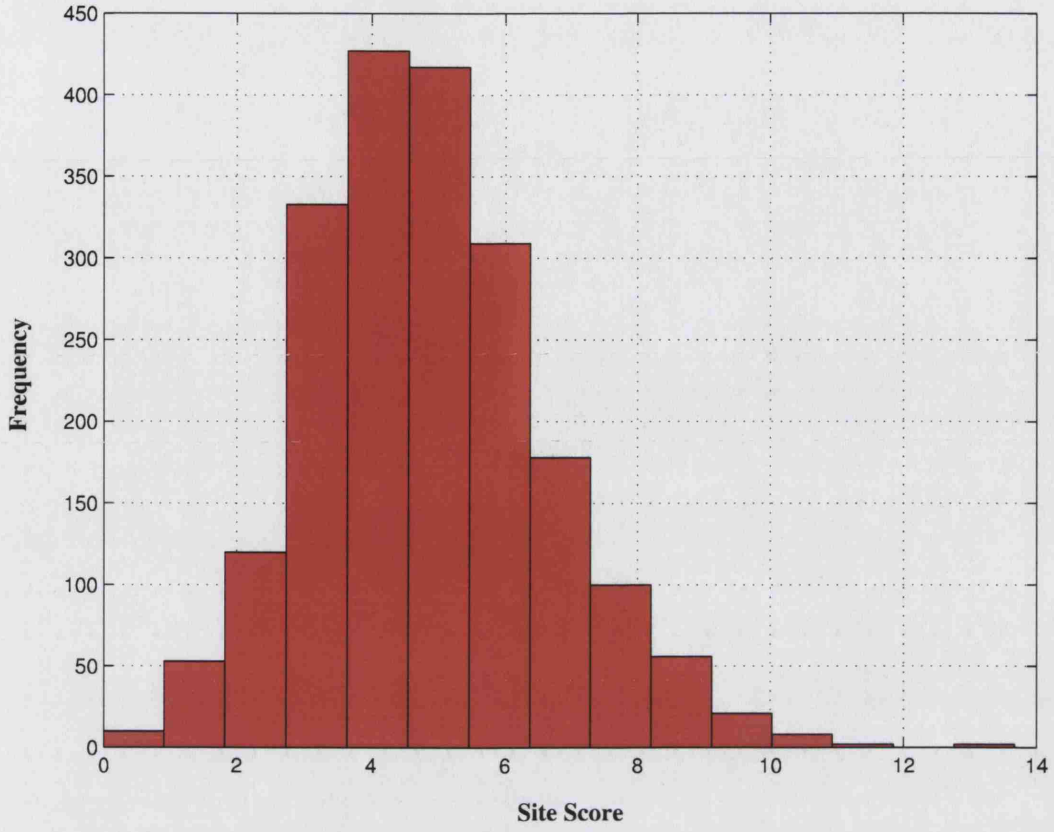


Figure 4.6: Distribution of DNA Site Score for genome models with unknown molecular function GO annotations

DNA (53%) at the 1% FPR threshold. This result provides further validation of the classification system and demonstrates effective DNA binding prediction in the GenTHREADER predicted structures.

Relaxing the site score threshold to 10 (2.7% FPR) resulted in 6 additional hits for DNA binding in the unknown function set as compared to the 1% FPR threshold. Table 4.6 provides a summary of the unknown function hits, along with the gene identifier, gene description as well as information from all three GO ontologies. The

hits include three proteins described as hypothetical ORF's and two annotations which are ambigous. For example, GI6323165 was annotated as being required for cell viability although no further information was available. Intrestingly, one hit (GI6323165) is a homolog of zinc finger protein which is annotated as DNA binding as well as unknown function in GO. Where cellular component information was available, all hits were annotated to be found in the nucleus.

Gene ID	Gene Description	Top Ranking Score	False Positive Rate (%)	Selectivity (%)	Gene Ontology Term		
					Molecular Function	*Cellular Component*	*Biological Process*
GI6323094	Hypothetical ORF	13.67	0.09	92.0	Unknown	Unknown	Unknown
GI6320973	66S pre-ribosomal particles, involved in 60S ribosomal subunit biogenesis	12.89	0.32	80.6	Unknown	Nucleus	Protein biosynthesis
GI6322689	Hypothetical Protein	11.52	1.06	69.2	Unknown	Nucleus	Cell growth
GI6323165	Zinc finger containing homolog of mammalian TIS11 gene	11.50	1.06	66.7	DNA binding, Unknown	Nucleus, Cytoplasm	Unknown
GI6321684	Severe Depolymerization of Actin	10.62	1.8	57.0	Unknown	Nucleus	Cytoskeleton
GI6321774	Hypothetical ORF	10.50	2.0	55.1	Unknown	Unknown	Unknown
GI6321264	Protein involved in bud-site selection; diploid mutants display a unipolar budding pattern instead of the wild-type bipolar pattern	10.42	2.1	54.6	Unknown	Nucleus	Bud site selection
GI6323080	Protein required for cell viability	10.02	2.7	49.1	Unknown	Nucleus	Unknown

Table 4.6: High scoring DNA binding predictions for *S. cerevisiea* proteins annotated with GO molecular function term 'unknown'.

4.4 Conclusions

Accurately identifying regions of a protein capable of interacting with DNA as well as discriminating between DNA binding function from non-DNA binding proteins is a particularly important area of research. The aims of the work presented here were to extend the site classification scheme, originally developed for identifying metal binding sites, to the automatic prediction of DNA binding proteins and their interacting residues. Accurate detection of such properties will not only be beneficial for genome annotation but could also aid targeting or even re-engineering of proteins through site-directed mutagenesis.

Combining structural and sequence features provided greater classification accuracy for identifying interfacing regions in proteins known to bind DNA. To represent the DNA interface residue in a more accurate manner, a simple extension of the method was developed. Predictions were grouped to surface patches of the protein, in a similar fashion to that described by Jones et al. (2003b). The benchmarking results demonstrate effective prediction accuracy and sensitivity using this approach. Of the dataset of 56 protein-DNA complexes, 35 proteins contained ≥ 7 true DNA interacting residues in the top ranked patch. Although the electrostatic patch approach in the Jones *et al.* method correctly identified 38/56 proteins using this criterion, the current findings are encouraging as several hits missed by the electrostatic method were correctly located. Further, the results demonstrate effective

extension of the MetSite method for DNA interface prediction and highlight that the information encoded within the PSI-BLAST PSSM's and low resolution structural features provides comparable results to atomic level electrostatic calculations.

The major disadvantage of the electrostatic approach is illustrated by the fact that large positive electrostatic patches are not unique to DNA binding proteins. The current method was therefore benchmarked on a dataset of DNA-binding proteins, and a decoy set of non-DNA binding proteins which include proteins that exhibit positive charged patches on their surface.

Another advantage of the system developed here is the applicability to any type of DNA binding protein. Classification by structural motif is accurate even when the neural network is cross-validated by structural motif. Therefore unlike other methods the DNA classification is not reliant on a specific DNA binding motif (such as a HTH).

Analysis on a set of proteins known to bind DNA but crystallised in uncomplexed form highlighted effective classification for these cases. This provides evidence that the method is not adversely affected by conformational changes which occur during protein-DNA interactions. However, due to the limited dataset of unbound proteins it is not possible to assess the significance of these predictions. Nonetheless, the correct prediction of a novel DNA binding protein in the unbound set is an encouraging result demonstrating the identification of a novel DNA binding motif.

Overall the results provide encouraging findings. DNA/non-DNA binding discrimination is achieved at an accuracy of 87.8% with high sensitivity 70.4% (38/54 true positives) at a reasonable false positive rate of 10% (25/250 false positives). However, the published results of Stawiski et al. (2003) showed better discrimination as compared to the current approach. In their study an accuracy of 92.1% was achieved (44 true positives and only 14 false positives) from the same set of proteins. An important consideration with the Stawiski approach, however, is the requirement of several features which are more reliant on highly resolved atomic positions (electrostatic and hydrogen bonding potential). Given the low-resolution approach taken, it is encouraging that the current method, which should be less dependant on structural quality, is comparable to the Stawiski approach.

To highlight the advantages for genome annotation, DNA binding predictions were benchmarked using structural models for the *S. cerevisiae* genome. The Gene Ontology (GO) annotations for DNA binding function, as well as descendants of this term, were used to label sequences providing 195 target sequences classed as DNA binding. Structural models and site predictions were generated using an updated version of the distributed GenTHREADER resource. 2037 protein sequences were used in the non-DNA binding set.

The cross-validated DNA classifier was able to accurately classify 90.1% of the protein models as DNA binding or non-DNA binding. Of the DNA binding pro-

teins this represented a sensitivity of 34.3% with a low false positive rate of 1%. Analysis of the transcription factor class revealed sensitivity to be 53% for these structural predictions. Classifications for the smaller GO classess however showed less sensitivity.

In conclusion, this study has shown that effective detection of DNA binding function is possible using PSSM and low-resolution structural features. Furthermore, the results are comparable to other published methods which require features more dependant on structural accuracy. The genome-wide analysis helps to demonstrate the advantage of the system. Prediction for DNA binding was shown to be possible for modelled proteins known to interact with DNA. In addition, several DNA binding predictions were also observed for predicted structures with unknown molecular functions. This suggests the approach should be extensible to other functionally important regions such as the interface between two interacting proteins or other larger ligand binding sites. Furthermore, given the role of native disorder in DNA binding proteins, it is likely that disorder predictions, from methods such as the DISOPRED system, may enhance the discrimination of DNA and non-DNA binding proteins as well as improve the correct detection of interfacing residues.

Chapter 5

Improving Fold Recognition Model Quality

5.1 Introduction

5.1.1 The Relationship Between Protein Structure and Function

The accurate and efficient prediction of protein three-dimensional structure given only the sequence has become one of the greatest challenges in modern biological sciences. The prospect of rapid elucidation of the structure of a target sequence has implications for understanding fundamental biological processes and interactions and is a goal which should not be underestimated.

An important aspect of both protein structure and function is the conservation of residue positions. Conserved residues may be placed in close proximity in the folded structure to form an active site or an interaction surface. In many cases the presence of a functional site often requires the specific arrangement of residues and atoms which are positioned by the overall fold of the protein. Therefore residues involved in maintaining structural features are also more likely to be under greater selective pressure.

The focus of the current chapter has been the application of the metal site classifiers, presented in Chapter 3, to assess the quality of structure predictions. This application is based upon the observation that conserved residues are often closely packed in the three-dimensional structure although they may be distant in

sequence. Therefore, in the correctly folded structure, residues linked to functional sites or involved in maintaining structural properties are more likely to be clustered in space. This is unlikely to be true for incorrect fold predictions.

The classification results from the MetSite analysis highlighted the important contribution of residue conservation and, to lesser degree structural features, in discriminating between sites and non-sites (section 3.3.1). Here we propose the features and site encoding used in the MetSite system could provide an effective means to quantify structural predictions.

5.1.2 Closing the Sequence-Structure Gap

Experimental routes for determining protein structure are by no means trivial whereas sequencing has now become an established technique carried out routinely. As a result the sequence repositories currently provide a greater representation of protein families as compared to structural repositories. However, protein structure is known to be more conserved as compared to sequence, estimates suggest that 15,000-20,000 GenBank sequence protein families can be grouped into approximately 1200 structural superfamilies (Orengo et al., 2003).

This highlights that proteins with identical folds can function very differently. The 'superfolds' demonstrate this principle: these folds consist of commonly occurring protein folds configurations (Russell et al., 1998). Analysis of 'superfold'

structures has revealed these proteins perform varying biological functions. Interestingly many 'superfold' proteins have been shown to contain preferred binding site locations known as 'supersites', such as those in TIM barrel and Rossman fold proteins (Russell et al., 1998; Orengo et al., 2003).

Bridging the gap between structure and sequence is the focus of numerous research efforts world-wide, below is a brief overview of the main strategies employed and challenges which remain.

5.1.3 Structural Genomics

The aim of structural genomics initiatives is the experimental determination of protein structures for every sequence-structure representative. It is hoped that such detailed structural information will allow elucidation of protein function. However, high throughput structural determination is both extremely costly and time consuming. Furthermore, sequence analysis tools and modelling techniques in cases where proteins share sequence identity $\geq$30% to known structure can usually yield confident structural assignments. The rapid rate of expansion of the protein databank (PDB), albeit less than that of sequence databanks, has certainly identified shortcomings in the structure to function area. This is exemplifed by the growing number of structures for which function has not been fully characterised.

The initial phase of the structural genomics initiatives came to completion in

2004 and although the go ahead has been given for the second phase several challenges have been identified. It is evident that target selection is ineffective, homologs to given targets maybe deposited during the course of structure determination resulting in significant costs and wastage. It is certainly clear a better management and tracking system needs to be in place to prevent such resources being wasted and focus on truly important targets.

Computational methods, however, offer the possibility of significantly improving the efficiency of structural genomics. Sadly the effective use of structural bioinformatics tools is not widespread, partly due to lack of awareness and also the need for specialist knowledge. Rapid computational approaches generally involve either homology modelling or fold recognition. Such methods certainly offer the potential of reducing the experimental burden or alternatively may be used to target novel folds. Both approaches could therefore increase the efficiency of experimental structural biologist and benefit the research community greatly.

5.1.4 Structure Prediction

Homology Modelling

A high degree of sequence similarity generally indicates a strong structural and functional relationship. Homology modelling methods such as the MODELLER program, developed by Sali et al. (1995), often provide good quality structural mod-

els if the structure of a close homolog is available. However, the requirement for homology is often restrictive. Assessment studies have shown that model quality falls rapidly as identity between target sequence and structural template sequence drops below 40% (Moult et al., 1997; Orengo et al., 2003). Consequently this approach is not always applicable, although the situation is likely to improve as the PDB coverage increases.

Fold Recognition

Fold recognition methods attempt to assign structural folds to target sequences in the absence of obvious sequence level similarities. This strategy relies on the fact that protein structure space is more redundant than both sequence and function. The target sequence is essentially mapped to a protein of known fold. Such approaches are obviously limited in the fact that they assume a target sequence adopts a known fold. However, fold recognition offers the potential of uncovering very distant evolutionary relationships which are undetectable by even the most sensitive sequence profile methods. Consequently, a great deal of research effort has been dedicated to correctly identifying the fold of a query sequence. The challenge is twofold; firstly to optimise the sequence to structure alignment, followed by discrimination of native-like predictions from decoy hits.

Ab initio

Methods which use only first principles (*ab initio*) to construct a model of the protein structure offer the possibility of predicting novel folds. These approaches tend to search the conformational space within which the protein chain may fold to locate the global, or lowest attainable, energy minimum. However a complete and viable solution to this problem is yet to be discovered. Nonetheless, methods which build models from libraries of super-secondary structure motifs, such as FRAGFOLD (Jones, 2001) and ROSETTA (Simons et al., 1999), have made some progress in this area. Unfortunately the computational burden of these approaches is still generally prohibitive for whole genome scale structure prediction.

5.1.5 Assessing Structure Prediction Methods

A quantitative comparison of structure prediction methods is not a trivial task. Individual methods must be benchmarked in an identical manner to allow unbiased assessments. The CASP (Moult et al., 2001), CAFASP(Fischer et al., 1999) and LiveBench(Rychlewski et al., 2003) initiatives are an ideal setting for such comparisons, providing a framework for blind predictions to take place. Such community wide assessments allow developments in the field to be measured and recognised as well as highlighting success and failures in given methods.

5.1.6 Measuring Model Quality

Traditionally an important problem in the structure prediction field has been an effective approach to measure the quality of a protein model. A common approach relies on calculating the root mean square deviation (RMSD) following an optimal superposition between the predicted protein and the native structure. However this approach is dependent on protein length and maybe adversely affected if a particular region is poorly predicted.

More appropriate alternatives have been developed as part of the on-going community wide assessment initiatives. The MaxSub (Siew et al., 2000) and LGScore (Cristobal et al., 2001) algorithms are two of the official methods used in LiveBench and CAFASP and allow the quality of a prediction to be rapidly and quantitatively determined.

5.1.7 Scoring Fold Recognition Models

An important challenge in fold recognition is the accurate and efficient discrimination of native or native-like protein models from decoy structures. Moreover, given a set of native-like models one would like to maximise the quality of the structural prediction. Learning based approaches have proved to be effective at tackling this problem, although scope for improvement still exists. Bjorn and Elofsson (Wallner and Elofsson, 2003) recently presented a neural network based method to improve

model quality. The approach involved training the neural network based on MaxSub and LGScore model quality measures and was shown to be effective at eliminating high scoring decoy predictions.

5.1.8 GenTHREADER

The GenTHREADER algorithm for fold recognition was one of the earliest approaches for rapid fold assignment (Jones, 1999; McGuffin and Jones, 2003). The method combines various parameters, including threading potentials originally reported by Jones (Jones et al., 1992), into a simple feed-forward neural network classifier producing two outputs. The original version of GenTHREADER was trained using binary targets (0,1) to signify a fold match as defined by CATH code. Confidence was estimated as the difference between the two output nodes. Successive versions of GenTHREADER (mGenTHREADER) were improved by seeding PSI-BLAST profiles using FSSP alignments and training to high and low Z scores produced by FSSP. Network outputs were then mapped to a p-value for a more robust confidence estimate.

GenTHREADER was recently used to construct the Genomic Threading Database by implementing a distributed version of the method (McGuffin et al., 2004a,b). This resource aims to provide genome-wide fold assignments. Currently over 200 genomes have been annotated, a task made feasible by the distributed computing

model. Genome-wide structural assignments within resources such as the GTD offer the potential to perform detailed analyses to understand protein interactions and functions. However, the improvement of the underlying fold recognition approach is an on-going challenge which is likely to improve the quality of genome annotations.

5.1.9 Chapter Overview

The aim of the research presented in this chapter is to determine the effectiveness of using the site detection methods, described in the preceding chapters, to improve the correct recognition of folds as well as the improvement of quality GenTHREADER predictions. The MetSite score is used in combination with a secondary structure element alignment score and model checking metric to develop a new neural network approach, nFOLD, to re-rank the initial predictions and improve the model quality of the top ranking solutions. Unlike the original GenTHREADER, training is performed to predict MaxSub score and therefore model quality. All neural networks tested show significant improvements over the unprocessed GenTHREADER predictions. Although in isolation the additional inputs do not improve model quality combining the scores resulted in statistically significant improvements.

5.2 Methods

5.2.1 Datasets

A fold library consisting of 3475 distinct protein chains was used to benchmark the nFOLD method, all proteins in this dataset share sequence identity <30% and FASTA E-value >0.01.

Benchmarking was assessed using two independent strategies; firstly cross validation was performed by randomly splitting the fold library dataset into five groups. This analysis was supplemented by validation on newly released structures from the LiveBench assessment of protein structure prediction methods. These structures are particularly convenient as they share no significant sequence similarity to known protein structures and often include structural genomics targets.

5.2.2 GenTHREADER

The GenTHREADER protocol combines six parameters into a neural network to provide a measure of similarity between a target sequence and template structure. The parameters are: sequence profile alignment score, number of aligned residues, length of target sequence, length of template sequence, pairwise energy and solvation energy sum. The pairwise and solvation potentials are calculated as described in the THREADER method (Jones et al., 1992). These potentials evaluate the quality

of the alignment between target sequence and template structure as well as scoring residue solvation. The potentials are derived from potentials of mean force using the inverse Boltzmann equation. The pair potential is based on the potentials originally described by Hendlich et al. (1990). For given atoms in a pair of residues ab with sequence separation k and distance s the pair energy is given by:

$$\Delta E_k^{ab} = RTln(1 + m_{ab}\sigma) - RT\ln(1 + m_{ab}\sigma\frac{f_k^{ab}(s)}{f_k(s)}) \tag{5.1}$$

where m_{ab} is the number of pairs ab observed with sequence separation k, σ is the weight for each observation, $f_k(s)$ is the frequency of occurrence of all residues at distance s, f_k^{ab} is the equivalent frequency for pair ab. RT is taken as 0.582 kcal/mol and k is defined as short (<11), medium ($11 \leq k \leq 22$) or long ($k > 22$) range. The pair potentials are calculated for the following atom pairs $C\beta \Rightarrow C\beta, C\beta \Rightarrow N, N \Rightarrow C\beta \Rightarrow O$ and finally $O \Rightarrow C\beta$. The solvation potentials are derived as follows:

$$\Delta E_{solv.}^{a}(r) = -RT\ln(\frac{f^a(r)}{f(r)}) \tag{5.2}$$

where r is the degree of residue burial, $f^a(r)$ is the frequency of occurrence of residue a with burial r and $f(r)$ is the frequency of occurrence of all residues with burial r.

5.2.3 Additional Inputs

In addition to the six original features used in GenTHREADER we evaluate three new features for improving model quality prediction: SSEA, MetSite and ModCheck scores.

Secondary Structure Alignment Score (SSEA) Score

SSEA scores were generated by the method implemented by McGuffin et al. (2001) based on the method originally reported by Przytycka et al. (1999). The method generates a score by aligning predicted and observed secondary structure elements using a dynamic programming algorithm based on that described by Needleman and Wunsch (1970).

SSEA scores are calculated using the observed secondary structures, recorded from DSSP output (reduced to either C, H or E), for the top 20 fold hits from GenTHREADER. The predicted secondary structure states, for the target sequence, were derived using the PSIPRED secondary structure prediction method (McGuffin et al., 2000). In order to prevent any bias, only cross-validated secondary structure assignments were taken. That is if a target protein sequence was found to produce a BLAST E-value ≤ 0.1 to any protein used to train PSIPRED then the corresponding weights were excluded for the secondary structure assignment.

MetSite Score

In the current study MetSite scores are assessed as a new means to quantify structural predictions which are generated from the established GenTHREADER fold recognition method.

MetSite predictions were derived by scanning every model from all target-template alignments with each of the six metal site classifiers which make up the MetSite method. The top ranking site based predictions are then extracted and the highest ranking MetSite prediction used to assess model quality. To prevent over-optimistic site predictions cross-validated weights are also used to generate MetSite scores as described above. A single MetSite score is obtained this way for every sequence-structure alignment.

An important consideration for the application of MetSite for fold detection is the site encoding scheme used in the system (described in section 3.2.2 and Figure 3.1). The approach defines several features of residues within a 3D space. Therefore for protein models which contain numerous gaps, due to inadequate sequence to structure mappings, obtaining a MetSite pattern is likely to be difficult. Such cases are therefore more likely to produce lower site scores as compared to models where conserved residues are placed within the same vicinity.

ModCheck Score

The model verification scheme ModCheck, developed by Jones and McGuffin (2003), was used to assess the quality of a model. ModCheck implements a simple approach in order to estimate the quality of a modelled protein. The input target sequence is randomly shuffled 1,000,000 times and the pairwise and solvation energy is calculated, as described above, between the random sequence mapped onto the protein fold being tested. The distribution of energy values is then used to calculate a Z-score which provides an estimate of the prediction quality.

5.2.4 Assessing Model Quality

Assessing the quality of a model protein is in itself a non-trivial problem. Several approaches have been described in the literature each with their own advantages and disadvantages. In the current study we have decided to use the MaxSub (Siew et al., 2000) algorithm to assess the quality of a predicted structure.

MaxSub is one of the official model quality checking programs used at CASP, CAFASP and LiveBench. The algorithm transforms and translates two structures to maximise the superposition of Cα atoms. The algorithm then determines the number of residues which fall within a user defined cut-off over the total number of residues in the experimental structure. This was set to 3.5Å for the present study in accordance to the assessment used on LiveBench and CAFASP.

The advantage of the algorithm is that it is extremely quick and produces a normalised score in the range 0-100 which represents the proportion of residues correctly located in the model structure. A score of 0 indicates no fold similarity whilst scores closer to 100 suggest stronger structural similarity. Scores between different template-target pairs can therefore be compared directly. The MaxSub score is rescaled between 0-1 in order to train the different neural network configurations.

5.2.5 Generating Structural Models

In order to scan query protein sequences for potential functional sites, structural models must first be generated. The initial top 20 GenTHREADER predictions are used to build the model structures which are then scanned using MetSite. In order to achieve this, the distributed version of GenTHREADER (McGuffin et al., 2004a,b) was used to generate the alignments and model structures. The MetSite classifiers were incorporated into the distributed system to automatically assign site scores for all the models (collaborative work with McGuffin, L.J).

The distributed system consists of 50 dual processor nodes (100 processors). This enabled the sequence-structure alignments using GenTHREADER to be performed within 3 days for the 3465 fold library dataset (67,982 sequence-structure pairs). The top 20 GenTHREADER predictions were used to build model proteins from which MetSite and ModCheck scores were calculated. The process of calculating

and extracting the data from the different sources is summarised in Figure 5.1.

5.2.6 Benchmarking Inputs

Several variations of the GenTHREADER method were generated and benchmarked in this analysis. In all cases the six inputs from the original threading potentials are included in training and the number of nodes in the hidden layer fixed to the number of nodes in the input layer. The neural networks produce a single output in the range 0-1. In addition to the six inputs the effect on fold detection was assessed using three additional features. Firstly, we include the top ranking site score derived from the set of MetSite classifiers. In addition the effect of SSEA and ModCheck scores are also evaluated.

Cross-validating MetSite and PSIPRED predictions

Secondary structure was predicted for all proteins using the PSIPRED secondary structure prediction method. In order to ensure fold predictions were stringently cross-validated, jack-knifed weights for both PSIPRED and MetSite were used. If any homology was detected between a target protein and any protein used to train either MetSite or PSIPRED then the corresponding set of weights was excluded.

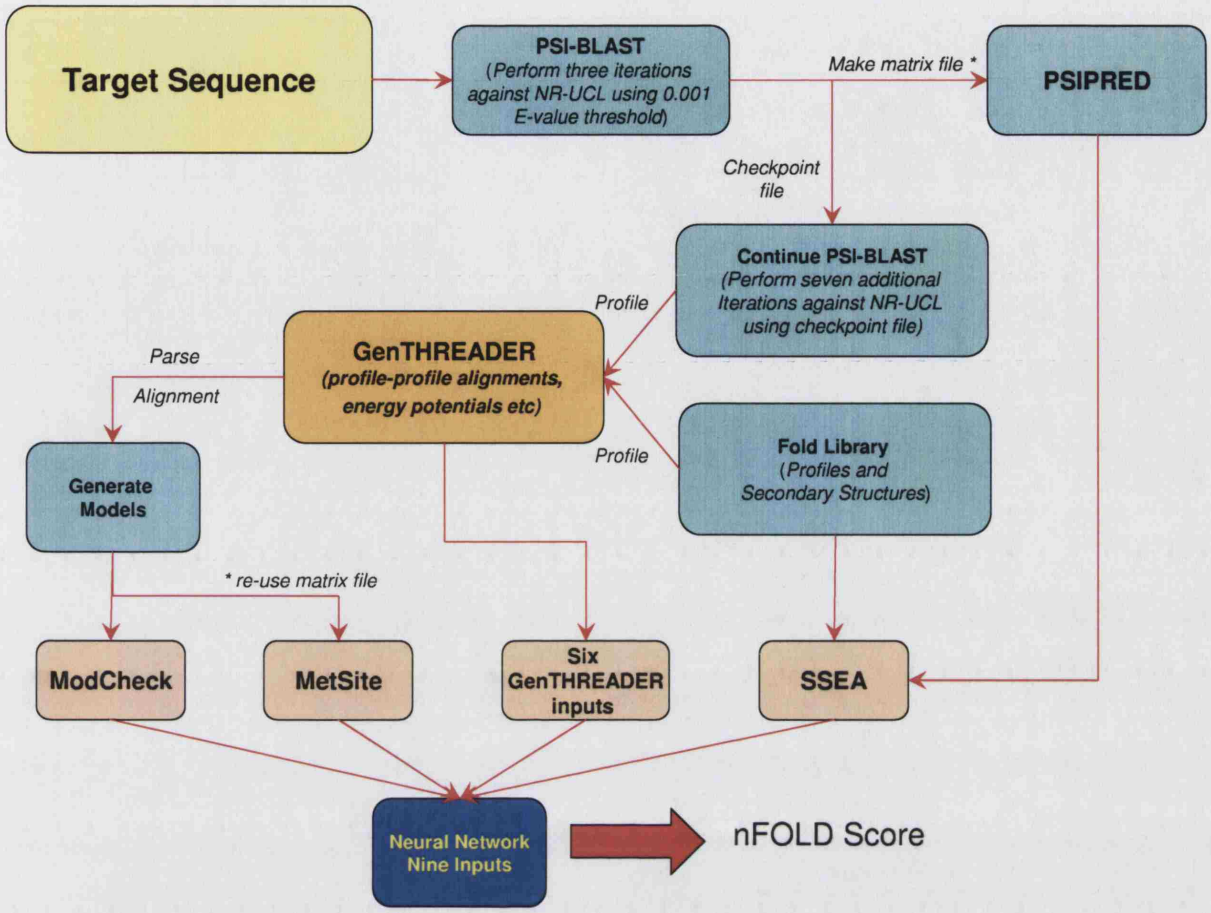


Figure 5.1: Diagram illustrating the process by which parameters are derived for the development of the nFOLD method. The six GenTHREADER parameters are extracted for the top 20 target-template alignments (ranked by E-value). ModCheck and MetSite scores are derived from structural models of these alignments. The SSEA score is derived from the alignment between the DSSP secondary structure states of the template structure and the secondary structure states of the target sequence, predicted by PSIPRED.

5.2.7 Network Structure and Training

The Matlab Neural Network toolbox v6 was used to set up and train several feed-forward neural networks. All networks were trained using the resilient back-propagation algorithm implemented within Matlab. The networks implemented all contained the same number of hidden nodes as the number of inputs. A three layer architecture (input, hidden, output) was used with tangent and logarithmic sigmoid transfer functions between the input-hidden and hidden-output layers respectively. Figure 5.2 summarises the networks used.

In order to optimise protein model quality we train all networks on observed MaxSub score. It should be noted that this is in fact a different approach to how the original GenTHREADER is trained. Therefore the implementation of GenTHREADER used here is in actual fact a simulation of the GenTHREADER inputs rather than the actual method and will be referred to as mGT6.

5.2.8 Normalising Score

SSEA produces an output in the range 0-1 and therefore was used directly as input for training. However, both ModCheck and MetSite scores were normalised using the sigmoid function below.

$$y = \frac{1}{1 + \exp^{(-ax)}} \tag{5.3}$$

where x is the raw input value, a an arbitary constant and y is the rescaled value.

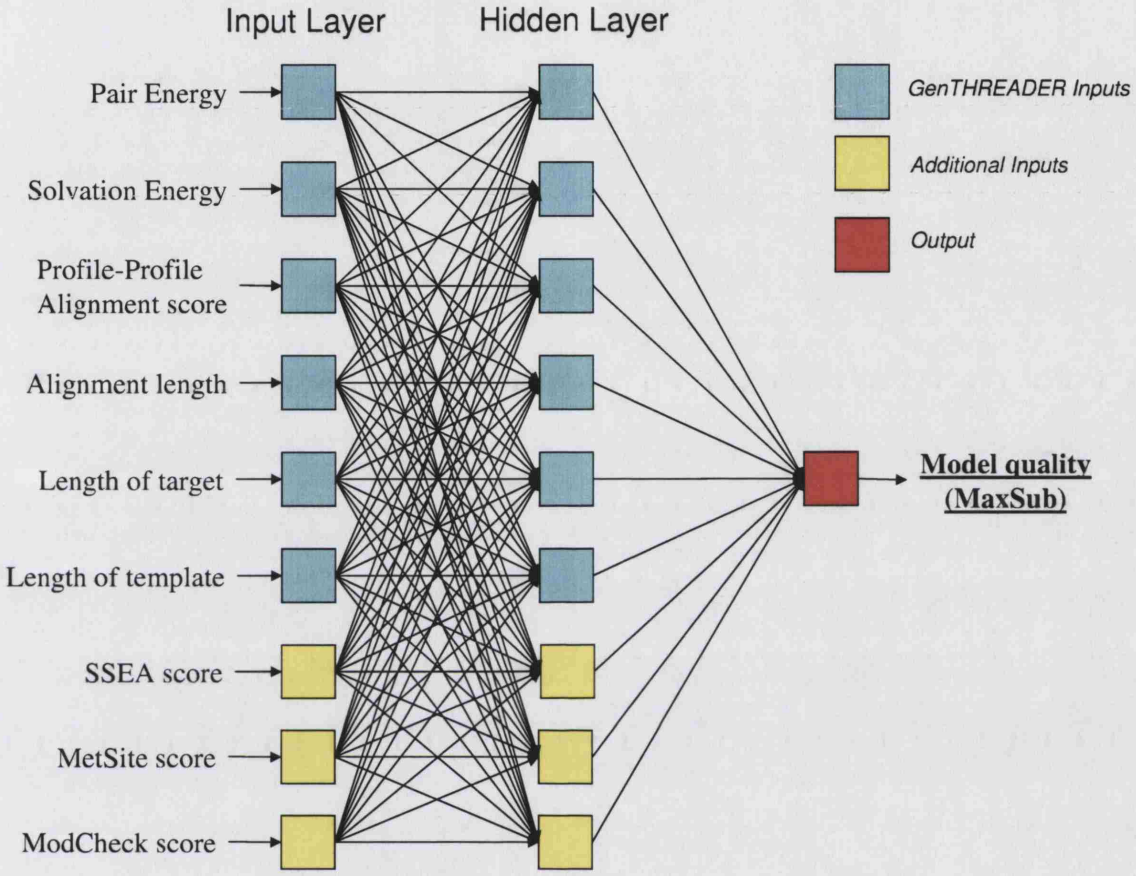


Figure 5.2: Diagram of neural network architecture used for training. Each of the inputs is specific to a given target-template alignment.

5.3 Results

5.3.1 Distribution of Method Score vs Model Quality

The distribution of scores for each of the three methods tested in this study were assessed against model quality. All 67,982 cross-validated target/template pairs, derived from the distributed system discussed above, were included in the analysis. The SSEA, MetSite and ModCheck scores were automatically extracted from every model generated along with MaxSub score. Target/template alignments producing a MaxSub score >0 are classified as true (positive) cases whereas MaxSub scores of 0 represents no fold similarity and therefore false (negative) cases. The score distributions, for each method, are plotted for both positive and negative cases in the following section.

Secondary Structure Element Alignment (SSEA) Score

The SSEA score distributions for the positive and negative cases differed strikingly from one another (Figure 5.3). The Figure clearly shows that SSEA scores for positive and negative cases can be easily distinguished. The Wilcoxon value, relating to the area under ROC curve (AROC), for this data was calculated as 0.75.

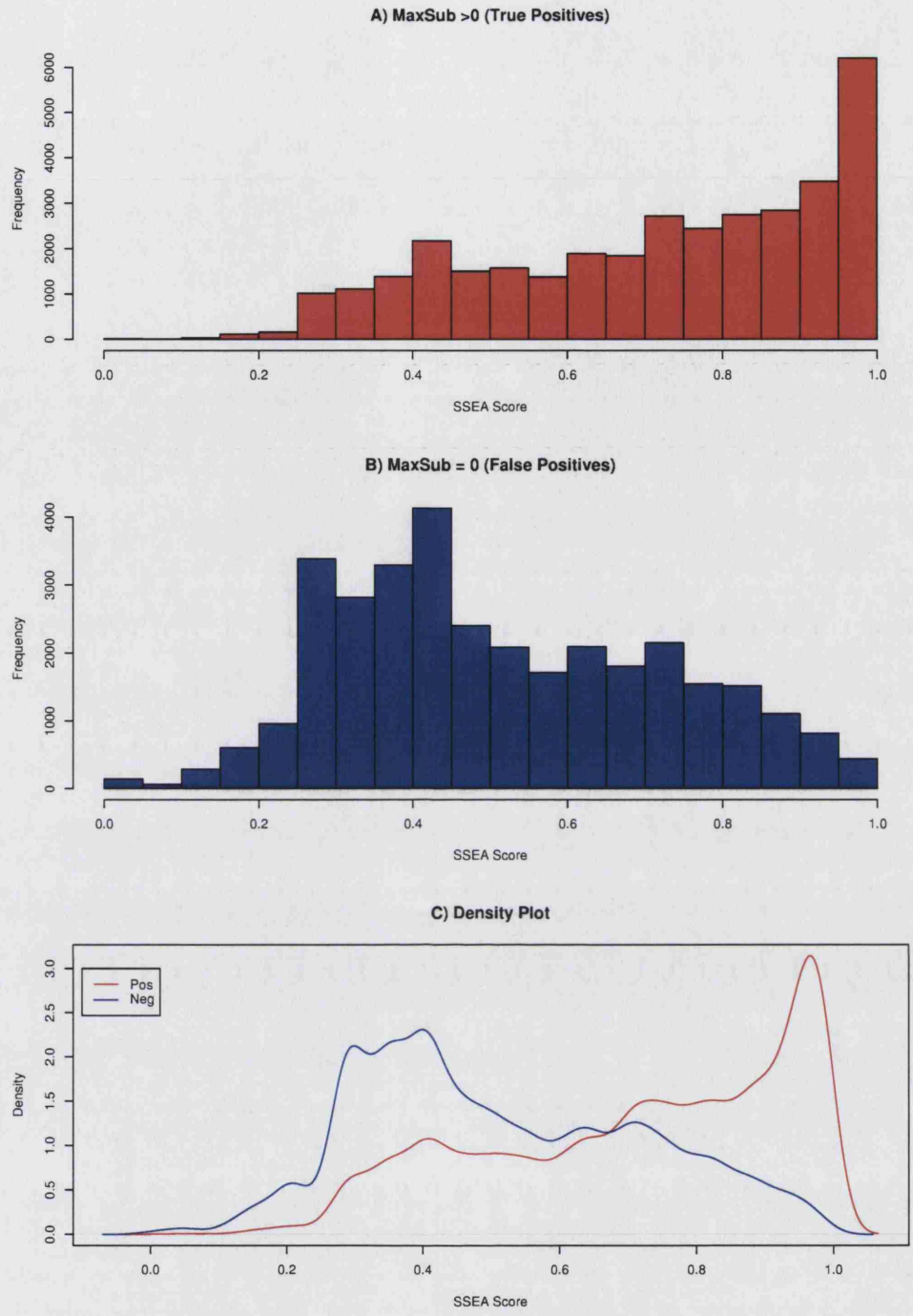


Figure 5.3: Distribution of SSEA Scores for A) true (MaxSub > 0) and B) false (MaxSub $= 0$) fold assignments for the fold library dataset. C) corresponding density plots.

MetSite Score

The distributions of top ranking MetSite scores for the positive and negative target/template pairs is presented in Figure 5.4. Visual inspection of the distributions suggests a greater overlap between the positive and negative sets as compared to the SSEA scores. Therefore the significance of the difference between the two MetSite distributions is less clear. However, the AROC value was determined to be 0.65 indicating that there is indeed the potential for MetSite to discriminate between true and false fold assignments.

ModCheck Score

The ModCheck score distributions are illustrated in Figure 5.5. The density plot in Figure 5.5c shows a distinct shift for positive cases (MaxSub >0) toward the higher range of ModCheck scores. Interestingly, the AROC value calculated to be 0.75 for the fold assignments.

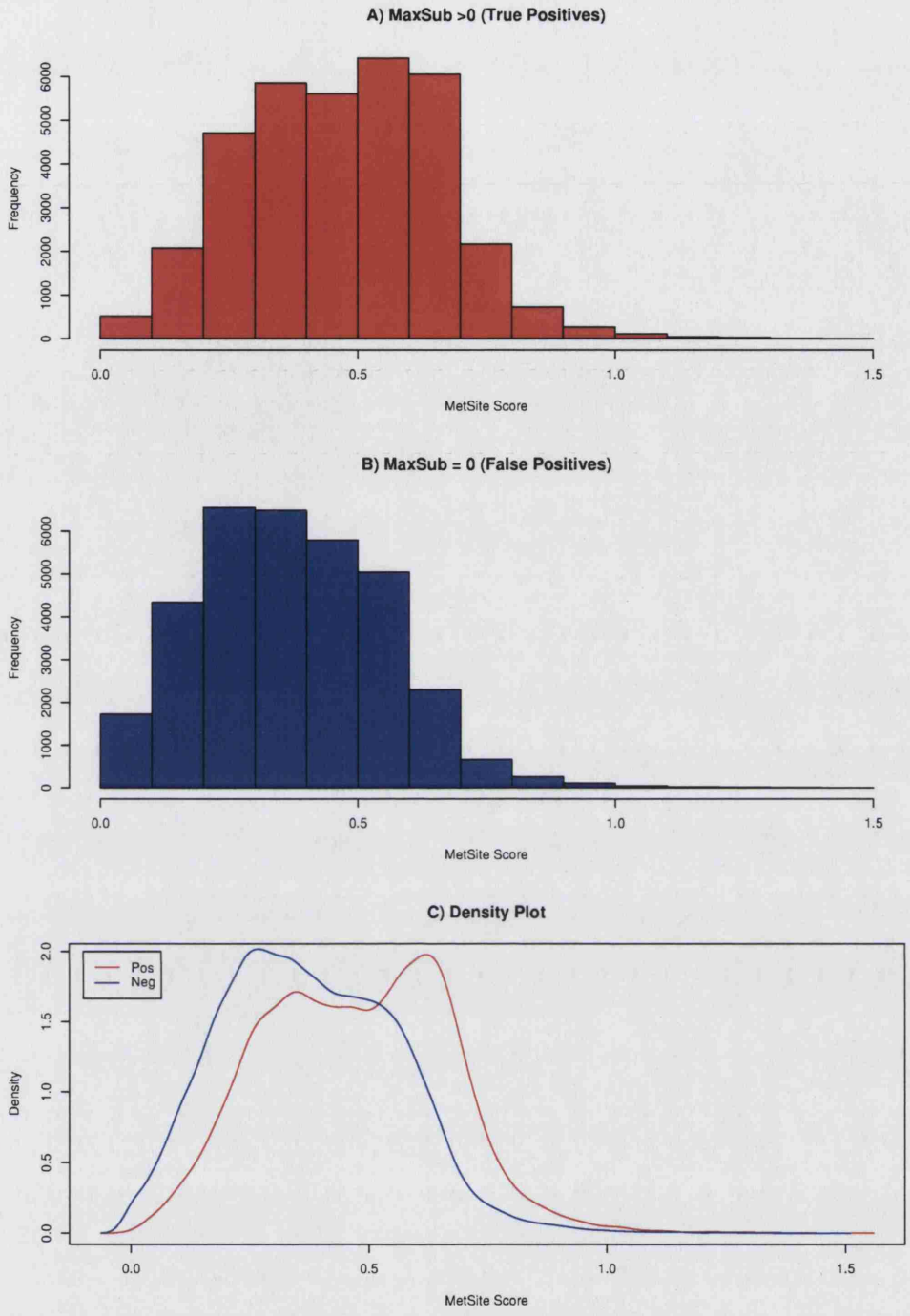


Figure 5.4: Distribution of MetSite Scores for true A) (MaxSub > 0) and B) false (MaxSub $= 0$) fold assignments for the fold library dataset. C) corresponding density plots.

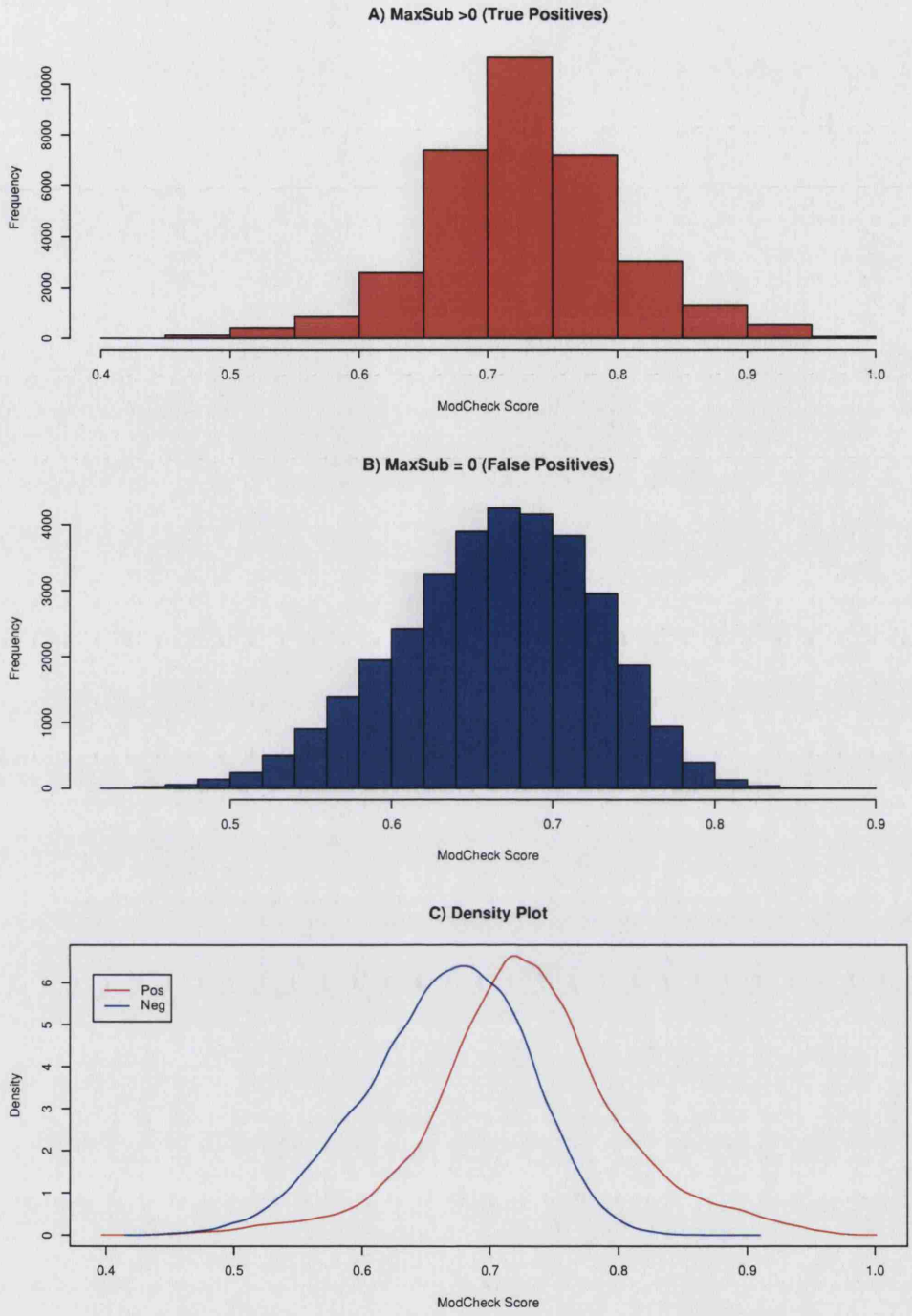


Figure 5.5: Distribution of ModCheck Scores for A) true (MaxSub > 0) and B) false (MaxSub = 0) fold assignments for the fold library dataset. C) corresponding density plots.

5.3.2 Ranking LiveBench Solutions

The 188 LiveBench-9 targets produced 1876 target-template alignments, taking the top ten predictions. The reason there were only 1876 pairs was due to the fact that there was a target in the dataset for which only six alignments were produced by GenTHREADER. The SSEA, ModCheck and top MetSite scores were extracted for each of these models and used to re-rank the model predictions. The MaxSub scores for the top ranking model solutions, using each of the three methods, are summarised in Table 5.1. Overall, re-ranking the predictions using the ModCheck score produced the fewest false hits (MaxSub =0). The total number of false hits for ModCheck was 66 followed closely by SSEA, with 67, and finally MetSite with 80.

Interestingly, there are several examples in the LiveBench re-ranking data which indicate the varying performance of the different methods. For example, there are 90 cases (47.9%) where the top SSEA ranked solution provides no useful model whereas either MetSite or ModCheck provide an adequate solution.

Target	Score Re-ranking				Target	Score Re-ranking			
ID	*SSEA*	*MetSite*	*ModCheck*	*Best MaxSub*	ID	*SSEA*	*MetSite*	*ModCheck*	*Best MaxSub*
1hl8A	0.117	0	0.23	0.244	1rkuA	0.258	0.311	0.287	0.462
1j1tA	0	0	0	0	1rlhA	0	0	0	0
1j26A	0	0	0	0.397	1rliA	0.474	0.517	0.555	0.579
1j3vA	0.376	0.477	0.244	0.477	1rniA	0	0	0	0
1j3wA	0.327	0.233	0	0.337	1ro8A	0.072	0	0	0.072
1lxyA	0.698	0.064	0.698	0.698	1rp4A	0.053	0	0	0.073
1m1lA	0	0	0	0	1rpuA	0	0	0	0
1n8nA	0.262	0.223	0	0.31	1rqiA	0.186	0	0.438	0.448
1nh1A	0	0	0	0.074	1rqpA	0.07	0	0.074	0.074
1nmoA	0.089	0.075	0	0.117	1rw2A	0	0.132	0.172	0.172
1nngA	0.463	0.535	0.463	0.555	1rwrA	0	0.082	0	0.127
1nnvA	0	0	0	0.279	1rwtA	0	0	0	0
1nrkA	0	0	0.463	0.463	1rxxA	0.439	0.439	0.412	0.439
1nxhA	0	0.19	0	0.19	1ry9A	0.327	0	0.384	0.384
1oapA	0.278	0.167	0.205	0.307	1ryaA	0.456	0.343	0.456	0.456
1ogkA	0	0	0	0.097	1rz3A	0.356	0.356	0.339	0.51
1oj5A	0.328	0	0.316	0.402	1s18A	0.075	0.072	0.075	0.087
1op4A	0	0	0.133	0.375	1s1iF	0	0	0.317	0.317
1p57A	0	0	0.39	0.39	1s2bA	0	0	0	0.121
1p8kZ	0	0.265	0.249	0.265	1s4nA	0.066	0.079	0	0.194
1p91A	0.24	0.325	0.304	0.325	1s5aA	0.399	0	0.267	0.479
1p97A	0.603	0	0.488	0.603	1s5lO	0	0	0	0
1p9eA	0.303	0.316	0.094	0.321	1s5lU	0	0	0.224	0.224
1p9hA	0.114	0	0	0.114	1s68A	0	0.223	0.223	0.273
1p9qC	0	0	0	0.093	1s79A	0.363	0.312	0.487	0.495
1paqA	0.128	0.159	0.159	0.159	1s7bA	0	0.236	0	0.236
1pc6A	0	0	0	0.156	1s7mA	0	0	0	0
1pfjA	0	0.229	0	0.255	1s7zA	0	0	0	0
1pm4A	0	0	0.174	0.174	1s9hA	0.237	0.095	0.163	0.237
1pmmA	0.472	0.336	0.353	0.472	1se9A	0.38	0.44	0.38	0.44
1psyA	0.356	0.241	0.336	0.356	1sedA	0	0	0	0.195
1pv5A	0	0	0	0	1sf8A	0	0	0	0
1pvmA	0.532	0	0.532	0.532	1sgoA	0	0	0.162	0.189
1q0dA	0.226	0.215	0.226	0.24	1si7A	0.058	0	0	0.066
1q67A	0.154	0	0.426	0.426	1sjwA	0.5	0	0.5	0.5
1q8rA	0.186	0	0	0.188	1skoA	0	0.512	0.179	0.512
1q9jA	0.206	0	0	0.206	1skoB	0.322	0.722	0.722	0.722
1q9uA	0	0	0.359	0.359	1souA	0	0.126	0.157	0.263
1qwrA	0.139	0	0.201	0.217	1sqhA	0.267	0.267	0.241	0.281
1qwyA	0.169	0	0.122	0.169	1sqsA	0.403	0.287	0.422	0.44
1qxmA	0	0	0	0.29	1squA	0.149	0.166	0	0.166
1qz4A	0	0.095	0.092	0.095	1sr4A	0.4	0	0.432	0.432
1r0uA	0.152	0	0.152	0.152	1sr4B	0.243	0.189	0	0.312
1r1dA	0.478	0.505	0.463	0.505	1sr4C	0.157	0.41	0.41	0.41
1r21A	0.642	0.541	0.642	0.642	1sskA	0	0	0	0
1r44A	0	0	0	0	1st0B	0.065	0.155	0	0.195
1r4vA	0.23	0.156	0	0.23	1suuA	0.145	0.137	0.182	0.182
1r4wA	0.119	0	0.343	0.343	1sv6A	0	0.255	0.445	0.445
1r57A	0.378	0.378	0.392	0.395	1szqA	0	0	0	0.049
1r5iD	0	0	0	0	1t16A	0.144	0.164	0.19	0.19
1r5jA	0.452	0.101	0.259	0.452	1t35A	0.499	0.499	0.206	0.499
1r5zA	0	0	0.074	0.074	1t5eA	0.246	0.246	0.246	0.246
1r61A	0.094	0.109	0.144	0.144	1t5jA	0.074	0	0	0.074
1r6fA	0.075	0.079	0.094	0.094	1t5yA	0.385	0.345	0.385	0.385
1r71A	0.205	0	0.276	0.276	1t62A	0	0	0	0
1r8iA	0.166	0.177	0.146	0.177	1t82A	0.54	0.448	0.656	0.656
1r9fA	0.939	0.939	0.939	0.939	1t8hA	0	0	0	0.098
1r9lA	0.147	0.147	0.226	0.226	1t9kA	0.084	0.067	0.067	0.192
1rcwA	0.489	0.55	0.489	0.55	1tc5A	0.595	0.595	0.586	0.595
1rfzA	0.135	0	0.115	0.141	1te5A	0.452	0.452	0.452	0.452
1rh5A	0.055	0	0	0.062	1tikA	0.321	0.344	0.564	0.564
1rhyA	0	0.119	0	0.119	1to3A	0.377	0.192	0.22	0.377
1ri1A	0.347	0.347	0.364	0.449	1to6A	0.061	0.057	0.07	0.071
1riqA	0.05	0	0	0.089	1ub9A	0.426	0.468	0.671	0.736
1rj1A	0.152	0	0.192	0.195	1uc2A	0.045	0	0.042	0.045

Target	Score Re-ranking				Target	Score Re-ranking			
ID	*SSEA*	*MetSite*	*ModCheck*	*Best MaxSub*	ID	*SSEA*	*MetSite*	*ModCheck*	*Best MaxSub*
1udmA	0.529	0.115	0.616	0.616	1v8cA	0	0	0	0.356
1uhwA	0.496	0.181	0	0.507	1v8oA	0.152	0	0	0.152
1ujtA	0.462	0.453	0.453	0.586	1v9dA	0	0.071	0	0.116
1ujxA	0.486	0.475	0.486	0.486	1vavA	0.432	0	0.432	0.432
1ul7A	0	0	0	0.2	1vddA	0.157	0.289	0.126	0.289
1ul9A	0.477	0.126	0.504	0.506	1vduA	0	0	0	0.182
1umhA	0.142	0.133	0	0.142	1vh4A	0	0	0	0.08
1upkA	0.072	0.123	0.234	0.234	1vheA	0.245	0.25	0.341	0.344
1usuB	0	0	0.181	0.181	1vhgA	0.552	0.287	0.616	0.616
1ut4A	0	0	0	0	1vhnA	0.31	0.41	0.384	0.41
1utwA	0.056	0.062	0.066	0.256	1vhoA	0.282	0.371	0.371	0.404
1uunA	0	0	0	0	1vhvA	0	0.092	0.097	0.348
1uw4B	0.296	0.097	0.296	0.405	1vi1A	0.064	0.105	0.267	0.267
1uw7A	0	0	0	0	1vi7A	0	0.104	0.276	0.276
1uwwA	0.141	0.12	0.141	0.361	1vizA	0.374	0.256	0.332	0.374
1ux6A	0	0	0	0	1vjfA	0.524	0.524	0.524	0.524
1uxoA	0.562	0.521	0.562	0.563	1vjgA	0.482	0.495	0.561	0.561
1uynX	0.126	0.169	0.198	0.198	1vjhA	0.486	0.176	0.176	0.486
1v04A	0.205	0.16	0.212	0.241	1vjnA	0.275	0.252	0.275	0.275
1v2bA	0.248	0	0	0.248	1vjuA	0	0	0	0.088
1v2yA	0.296	0.391	0.475	0.475	1vjxA	0.63	0.675	0.586	0.687
1v32A	0	0.591	0.591	0.591	1vk0A	0	0.259	0.325	0.357
1v4eA	0.125	0.08	0.482	0.482	1vk5A	0	0	0	0
1v5mA	0.499	0.398	0.516	0.516	1vk9A	0	0	0	0
1v5oA	0.608	0.499	0.48	0.608	1vkbA	0	0	0	0
1v5pA	0.474	0.403	0.472	0.564	1vkfA	0.444	0.212	0.444	0.473
1v63A	0.353	0.384	0.435	0.435	1vkhA	0.34	0.345	0.3	0.392
1v74A	0	0	0	0	1vknA	0.237	0.155	0.345	0.345
1v7mV	0.276	0.354	0.154	0.354	1w0bA	0	0.216	0.281	0.281

Table 5.1: Re-ranking of LiveBench-9 fold assignments. The top ranking MaxSub score is presented for each of the 188 LiveBench targets using SSEA, MetSite or ModCheck score to rank solutions. The best MaxSub represents the maximum achievable MaxSub score in the top ten solutions.

Comparing Rankings

The above results were analysed to determine the frequency of cases where ranking by one method outperformed another method. The results presented in Table 5.2 provide an interesting insight into the performance of each method for detecting good quality models. At higher MaxSub thresholds, the gap between the ranking methods widens. A MaxSub score $\geq$0.2 was achieved for 34.6% of the the ModCheck ranked solutions. The SSEA and MetSite re-ranking only achieved this quality of predictions for 30.3% and 23.4% respectively. This order of performance is reproduced in the cumulative MaxSub scores for each approach, calculated as 3041.2, 3615.5 and 4064 for MetSite, SSEA and ModCheck reranking respectively.

Method	*Best*	Improvement(>0)			Improvement(>0.2)		
Score	*MaxSub*	*SSEA*	*ModCheck*	*MetSite*	*SSEA*	*ModCheck*	*MetSite*
SSEA	59	-	40	75	-	10	20
ModCheck	77	69	-	87	21	-	28
MetSite	50	54	53	-	11	8	-

Table 5.2: Summary of re-ranking analysis for LiveBench-9 dataset. Best MaxSub represents the number of cases for which the best possible MaxSub score was attained. The improvements relate to the number of cases where the MaxSub score in row i is greater than 0 or 0.2 than the method in column j.

5.3.3 Benchmarking Neural Network Inputs

The ongoing blind assessments of structure predictions methods, such as CASP, CAFASP and LiveBench highlight the shortcomings of fold recognition methods. It is not only important to maximise the quality of the predicted model but discriminate native-like hits from decoy predictions. This is effectively a ranking problem as quite often the fold recognition method ranks better solutions further down the list.

The 3465 fold library protein structures produced 67,982 target-template alignments. These were used to perform five-fold cross validation experiments in order to assess the influence of the new parameters for correct fold detection and improved model quality.

5.3.4 Statistical Significance of Top Ranking Predictions

The first question we address is the effect of the different inputs on the top ranking prediction for each individual target. This is achieved by ranking the cross-validated results by neural network outputs and extracting the top ranking network output for a given target along with the corresponding MaxSub score for the associated model-template pair.

In order to determine the statistical significance of the fold recognition results, paired Wilcoxon signed rank tests were performed for the MaxSub scores of the top

ranked solution for each network comparison (Table 5.3). A non-parametric test is used due to the fact that MaxSub score distributions of the top hits are not normal.

Neural Network	Wilcoxon Signed Rank Test			
	mGT6	+MC	+SSEA	+MS
mGT6	-	0.95	0.96	0.86
+MC	0.05	-	0.61	0.36
+SSEA	0.04	0.39	-	0.22
+MS	0.13	0.64	0.78	-
+MC_MS	5.1×10^{-4}	4.8×10^{-4}	7.7×10^{-4}	9.9×10^{-5}
+SSEA_MS	0.217	0.83	0.89	0.68
+SSEA_MC	1.7×10^{-4}	0.039	0.037	0.007
nFOLD	5.2×10^{-5}	6.4×10^{-3}	9.89×10^{-3}	2.1×10^{-3}
	+MC_MS	+MS_SSEA	+MC_SSEA	nFOLD
mGT6	0.99	0.57	1	1
+MC	1	0.17	0.96	0.99
+SSEA	1	1	0.903	0.99
+MS	1	0.312	0.993	0.99
+MC_MS	-	0.069	6.8×10^{-4}	0.21
+SSEA_MS	1	-	1	1
+SSEA_MC	0.931	5.1×10^{-4}	-	0.56
nFOLD	0.79	5.6×10^{-4}	0.44	-

Table 5.3: Calculated p-values for paired one-sided Wilcoxon Signed rank tests. The alternative hypothesis states that the top ranking MaxSub score for the method in row i is greater than that of the method in column j. Networks: mGT6 (six inputs from GenTHREADER), +MC (ModCheck), +MS (MetSite), +SSEA (secondary structure element alignment).

The Wilcoxon tests provide some interesting findings. It is clearly seen that the improvements observed by individually adding SSEA or ModCheck are only border-

line significant whilst adding MetSite alone did not significantly improve the quality of the top predictions. However, incorporating MetSite, SSEA and ModCheck in combination with the six mGT6 inputs provides the most significant improvement (p-value = 5.17×10^{-5}). The addition of ModCheck and MetSite provides a more significant improvement as compared to adding SSEA with ModCheck. Incorporating MetSite and SSEA on the other hand showed no significant improvement in model scores.

Although the Wilcoxon test provides a statistical framework within which to identify improved scores, the approach is unable to provide magnitude of the improvement in terms of model quality. This highlights a general problem with benchmarking structural predictions, however, the results do indicate a statistically significant improvement by introducing the additional inputs.

5.3.5 Assessing Model Quality

The MaxSub score for the top ranked predictions from the cross-validated results were compared across the different networks. The difference between the top MaxSub score from the mGT6 network as compared to the networks incorporating the additional inputs were calculated. Table 5.4 summarises the number of targets at varying levels of improvement measured as the difference in the MaxSub score of the top hit as compared to the top hit of mGT6.

The results in Table 5.4 indicate that nFOLD is able to improve the MaxSub score of the top prediction for 14.9% of target sequences as compared to mGT6. Moreover, 12.9% (68) of these improvements are greater than 0.2. Interestingly, the performance rankings change at different levels of improvements for the remaining networks trained with the addition of only a single score.

The inclusion of ModCheck results in more protein models with modest improvements in model quality (increases of up to 0.1 MaxSub score). However, adding only SSEA scores resulted in five more proteins predicted with MaxSub $\geq$0.3. This suggests that although ModCheck improves model quality more consistently, the addition of SSEA can improve the quality for a few predictions by a larger amount.

	Difference in MaxSub Score					
Network	< 0	0	>0	$\geq$0.1	$\geq$0.2	$\geq$0.3
nFOLD	141	2528	517	185	68	26
+SSEA	152	2677	406	145	58	27
+MetSite	283	2728	380	121	43	19
+ModCheck	191	2560	475	162	55	22

Table 5.4: Comparison of top ranking MaxSub scores of the improved networks versus the mGT6 method. The table shows the number of proteins for which a given network produces no difference in MaxSub score or improvements of at least 0.1, 0.2 or 0.3.

A top ranking prediction which has a MaxSub of 0 is a false positive. The cross-validated results were scanned to determine the number of occasions on which the

different networks failed to produce any useful structural information. The mGT6 cross-validated results produced 760 hits as the top rank with a MaxSub score of 0. The nFOLD approach obtained MaxSub score >0 for 73 of these 760 targets, giving an improvement of 9.6%.

5.3.6 Cumulative MaxSub

The sum of MaxSub scores for a set of predictions provides a measure of the quality of an ensemble of model proteins (McGuffin and Jones, 2003; Kelley et al., 2000). The cumulative MaxSub score was calculated for the top ranking model solutions from each neural network configuration. According to this metric, nFOLD outperformed all other networks with a score of 125,714. Inclusion of SSEA was observed to produce a sum score of 125,005 followed by MetSite with 124,671 and then ModCheck 124,610. Training using only the six original inputs produce a sum of MaxSub for the top-hits of 124,272.

5.3.7 Model Quality vs Error

The plots in Figure 5.6 shows the cumulative MaxSub score against false positives for the fold library analysis. We use a similar definition as described by Kelley et al. (2000) where a true similarity is counted only once for each target. Therefore only the top ranking correct solution contributes to the total MaxSub score to ensure

multiple similarities for a given target do not influence the assignments. This type of plot allows model quality of the predictions to be compared at equivalent numbers of false positives.

The cumulative MaxSub scores are consistent with the Wilcoxon analysis. The nFOLD network can be seen to produce a higher MaxSub total at equivalent false positives compared to the other networks. Combining MetSite and ModCheck provides the second best improvement followed by the ModCheck and SSEA combined network. The mGT6 network consistently obtains lower MaxSub total scores demonstrating the improvement in model quality using the additional inputs.

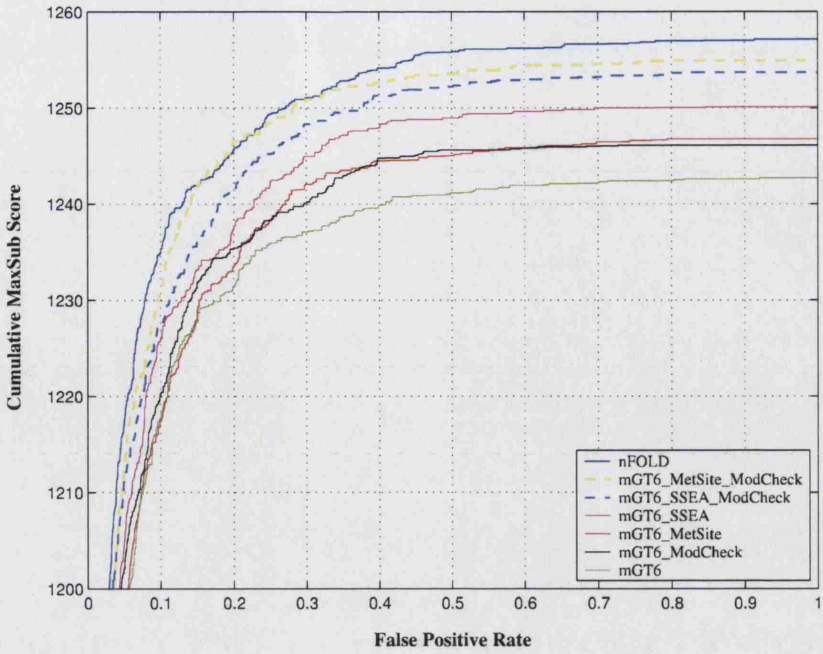


Figure 5.6: Cumulative MaxSub vs False Positive Rate for fold library cross-validation classification using the six GenTHREADER parameters in combination with MetSite, SSEA and ModCheck. False positives are defined as model proteins producing a MaxSub score of 0.

5.3.8 Assessment on LiveBench-9 Targets

Figure 5.7 shows the ROC plot for nFOLD on the LiveBench-9 dataset. For comparison, classification using the mGT6 network as well as the unprocessed results produced by GenTHREADER (original GenTHREADER) are presented. In addition, classification results obtained by ranking the solutions by only MetSite, SSEA or ModCheck scores (naive classifiers) are also included. True positives are taken to be any model protein with a MaxSub $\geq$0.3 as compared to its native structure.

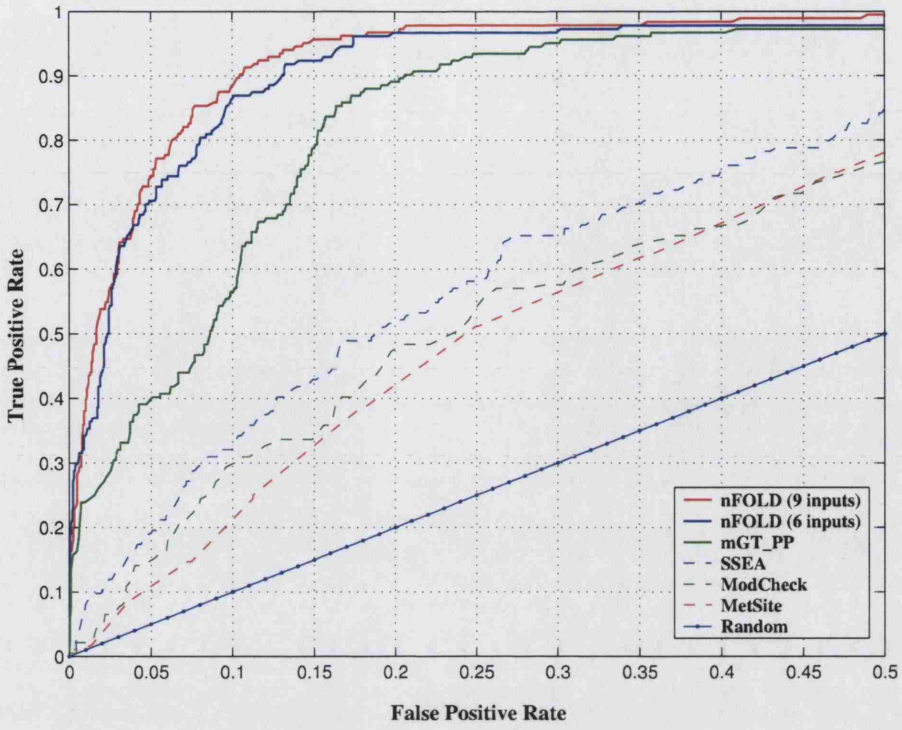


Figure 5.7: ROC plot for LiveBench-9 dataset using nFOLD, mGT6 or GenTHREADER neural networks or scoring by SSEA, ModCheck or MetSite scores.

Figure 5.7 highlights the marginal improvement the nFOLD classifier has over mGT6. However, for the LB9 dataset the improvement on the top hits is not statistically significant. The improvements over the unprocessed GenTHREADER results on the other hand are much more significant, at a fixed false positive rate of 5% nFOLD and mGT6 correctly identify 73% and 71% of models respectively with MaxSub $\geq$0.3 as compared to only 40% for GenTHREADER. This suggests that the majority of the improvement over the GenTHREADER predictions is due to

training on MaxSub score, whilst the additional inputs only have a moderate affect on classification.

5.3.9 Comparing network output to model quality

It is important that any method which attempts to rank solutions, has a reliable and quantitative score. Discrimination is not only required between native and decoy structures but also between near native structures. Figure 5.8 demonstrates the correlation between nFOLD and GenTHREADER scores against model quality as measured by MaxSub score. The correlation coefficient for nFOLD against MaxSub and GenTHREADER against MaxSub was calculated as 0.65 and 0.56 respectively for all 1876 target-template pairs from the LiveBench-9 dataset.

Interestingly, the distribution of GenTHREADER network outputs forms two distinct clusters, one at network output of 0.5 and another closer to 1. The MaxSub values for GenTHREADER, even at high network outputs, span a significant range, predominantly between 0.1 and 0.6. Clearly, in practice, this would suggest that a high GenTHREADER network output is not a reliable estimator of model quality. Conversely, the stronger correlation between nFOLD network output and MaxSub score provides a better confidence for predictions, as illustrated by the distinct cluster of predictions at low network outputs which are also scored poorly by MaxSub. This result is due to the nFOLD (and mGT6) being trained to continuous Max-

Sub score, as opposed to GenTHREADER which is trained to optimise distinction between high or low Z-scores from FSSP.

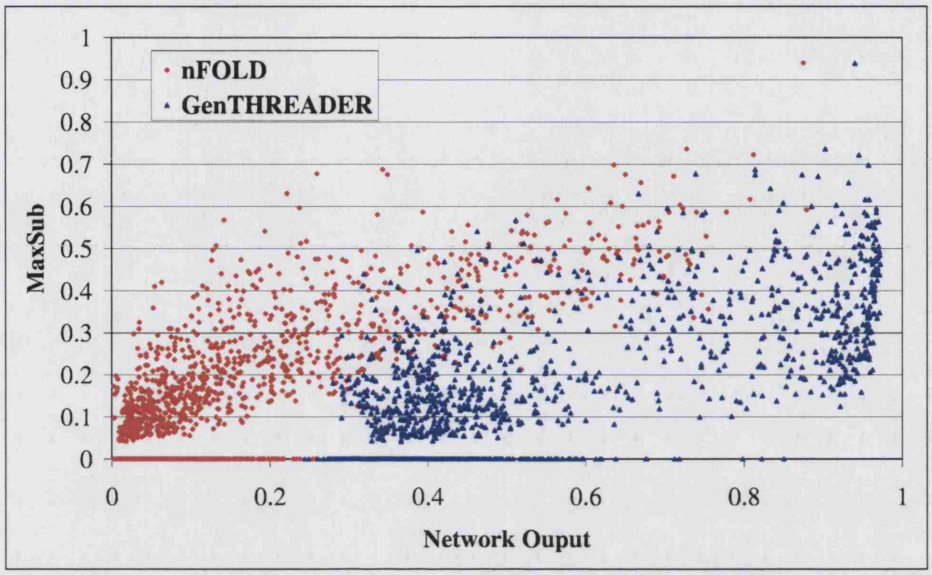


Figure 5.8: Network output against Maxsub score for all target-template pairs from the LiveBench-9 dataset.

5.3.10 Performance of nFOLD on LiveBench-9

Newly released structures used in LiveBench offer a convenient dataset for validation and comparison of different methods. These structures did not share any significant sequence similarity to known protein structures and allow different fold recognition servers to be compared directly. This comparison and ranking task is performed on

an ongoing basis by the LiveBench server.

The nFOLD method was implemented as a web-server (collaborative work with McGuffin, L.J) where users can submit a target sequence. Structural coordinates, in PDB format, of the top ranking prediction are subsequently emailed back to the user.

The nFOLD server was submitted for assessment on LiveBench to allow independent comparison with other structure predictions methods. The submission of nFOLD coincided with the the last month of the LiveBench-9 experiment (June 2004). The results provide an encouraging assessment of the method on the June targets, nFOLD was ranked as the third best fold recognition server as compared to non-meta server methods. This was two rankings above GenTHREADER.

5.4 Discussion

The focus of the work presented in this chapter has been to improve the detection of superior model quality structure predictions from a set of decoy and native-like structures produced by GenTHREADER. Previous work has shown that obtaining the overall fold of the protein, or parts of it, is generally feasible. However, an important task in rapid genome-wide fold recognition is not only to predict the correct fold but also to maximise the quality of the predicted model whilst providing

a reliability measure for the assignment.

The MetSite method was shown previously (chapter three) to be effective at locating metal site regions in low-moderate quality predicted structures (Sodhi et al., 2004). Here we propose that metal site regions, like other functionally and structurally important regions in protein structures, are more likley to be associated with correctly folded proteins, as compared to decoy structures. Such features include the close proximity of conserved residues, as well as similarities between packing and site environments.

The secondary structure element alignment (SSEA) score and a model checking algorithm (ModCheck) have been shown to improve structural predictions (McGuffin and Jones, 2003; Jones and McGuffin, 2003). We initially tested the ability of MetSite, SSEA and ModCheck to distinguish correctly predicted structures from a set of decoy predictions individually. The distribution of method score for true and false protein models revealed that SSEA scores provide the greatest discrimination. Figure 5.3 shows the distribution of SSEA scores for true fold predictions were strikingly different to the false cases. The result suggests that SSEA scores $\geq$0.7 are likely to represent true structural similarities. The score distributions, from each of the three methods, were shown to be significantly higher for true protein models as compared to false models at a confidence level of >99.9%.

Interestingly, ranking structure predictions using the scores from the different

methods revealed that ModCheck produced the least number of false predictions as the top hit. Furthermore, ModCheck obtained the best solution, given the original top 10 GenTHREADER predictions, more often. The results show that 34.6% of the ModCheck top ranked solutions produced MaxSub scores $\geq$0.3 as compared to 30.3% and 23.4% for SSEA and MetSite. However, Table 5.1 reveals some exceptions. Although overall ModCheck outperforms the other methods at obtaining the best possible MaxSub score, as might be expected, there are a number of cases where either MetSite or SSEA obtain significantly better scores. Notably, there are three cases where the MetSite top ranked solution is the best possible solution whilst re-ranking using SSEA and ModCheck produced hits with no structural similarity (MaxSub=0). This was compared to the random chance of obtaining such solutions by taking the total number of solutions greater than MaxSub of 0 (5 cases) over all 30 predictions for these cases (16.7%).

Given the varied performance of the different score systems, it was decided to combine the metrics with the original parameters used in the GenTHREADER method. Several different networks were evaluated and assessed using a similar strategy to that used in LiveBench. Training for all the networks was performed to actual MaxSub score target values in order to provide a better estimate of model quality.

Figure 5.6 shows the cross-validation results for a fold library, consisting of 3456

proteins, and highlights improvements in model quality as measured by the cumulative MaxSub score. The greatest improvement is demonstrated in a new method, nFOLD, in which the six inputs implemented in the original GenTHREADER approach are combined with SSEA, ModCheck and MetSite scores. The fold library cross-validation results demonstrate that the improvements in model quality, observed using the additional inputs, are statistically significant. Furthermore, structure predictions for a set of newly released structure from LiveBench-9 reveal a much greater magnitude of improvement for both the nFOLD method and mGT6 (GenTHREADER trained to MaxSub values) as compared to the original GenTHREADER.

Given the performance of the nFOLD approach, the method was automated and made accessible as a web-based server. Effective comparison and benchmarking against third party structure prediction methods is complicated by various technical problems, such as choice of fold library and cross-validation conflicts. Therefore the nFOLD server was submitted to LiveBench for continuous and automatic assessment. Due to the method being posted during the course of LiveBench-9 experiment several targets were not assessed, however results from the June 2004 target list (last month of LiveBench-9) provide some encouraging findings. The cumulative MaxSub scores for the top and best predictions for a selection of the top independent, non meta-server methods on LiveBench-9 server were retrieved. The results show the

nFOLD method performs extremely well on this blind assessment, ranking third based on top predictions and second on best prediction. It should be noted that as the nFOLD predictions are based on the underlying alignments generated by GenTHREADER the performance of the two approaches will be correlated. However, the benchmarking results and blind assessment on LiveBench indicate that nFOLD consistently outperforms GenTHREADER.

Although MetSite predictions, SSEA or ModCheck did not significantly improve model quality in isolation, combining all three the scores did. Interestingly, the combination of MetSite and ModCheck provided the second most significant improvement followed by combining ModCheck with SSEA. However, the MetSite and SSEA combined networks did not show improvements in model quality. A possible reason for this may be due to the fact that MetSite generally identifies sites at the ends of helices and loop regions due to the greater occurrence of metal ions in such locations.

The MetSite role in improving model quality is likely to be linked to features of a correctly folded protein. Regions where conserved residues are closer in space are more likely to represent correctly folded proteins as well as representing putative metal or perhaps other similar types of sites. Of course, the degree of conservation and nature of residues in close proximity will affect the results but such regions may come about through the structural conservation of interactions. Another con-

sideration is the site encoding scheme used in MetSite. If a protein model is poor it is likely to contain numerous gaps, obtaining a site pattern for such models will therefore be more restricted and therefore more likely to give a lower site score.

An important consideration for genome-wide structural annotation is related to the reliability of assignments. Figure 5.8 shows us that nFOLD predictions are more correlated to model quality as compared to GenTHREADER. This is perhaps not surprising as the original GenTHREADER method was not developed to optimise model quality directly but rather to discriminate between true and false protein folds. Therefore the results suggest that although the added inputs are more likely to provide a better quality model prediction, the majority of the improvement is actually due to training directly on observed MaxSub scores.

Bibliography

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., Lipman, D. J., 1990. Basic local alignment search tool. J Mol Biol 215 (3), 403–410.

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., Lipman, D. J., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25 (17), 3389–3402.

Andreeva, A., Howorth, D., Brenner, S. E., Hubbard, T. J. P., Chothia, C., Murzin, A. G., 2004. SCOP database in 2004: refinements integrate structure and sequence family data. Nucleic Acids Res 32 Database issue, 226–229.

Apweiler, R., Bairoch, A., Wu, C. H., Barker, W. C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, M. J., Natale, D. A., O'Donovan, C., Redaschi, N., Yeh, L.-S. L., 2004. UniProt: the Universal Protein knowledgebase. Nucleic Acids Res 32 Database issue, 115–119.

Artymiuk, P. J., Poirrette, A. R., Grindley, H. M., Rice, D. W., Willett, P., 1994.

A graph-theoretic approach to the identification of three-dimensional patterns of amino acid side-chains in protein structures. J Mol Biol 243 (2), 327–344.

Attwood, T. K., Beck, M. E., Bleasby, A. J., Degtyarenko, K., Michie, A. D., Parry-Smith, D. J., 1997. Novel developments with the PRINTS protein fingerprint database. Nucleic Acids Res 25 (1), 212–217.

Attwood, T. K., Croning, M. D., Flower, D. R., Lewis, A. P., Mabey, J. E., Scordis, P., Selley, J. N., Wright, W., 2000. PRINTS-S: the database formerly known as PRINTS. Nucleic Acids Res 28 (1), 225–227.

Bairoch, A., 1994. The ENZYME data bank. Nucleic Acids Res 22 (17), 3626–3627.

Bairoch, A., Apweiler, R., 1997. The SWISS-PROT protein sequence database: its relevance to human molecular medical research. J Mol Med 75 (5), 312–316.

Barker, J. A., Thornton, J. M., 2003. An algorithm for constraint-based structural template matching: application to 3D templates with statistical analysis. Bioinformatics 19 (13), 1644–1649.

Bateman, A., Birney, E., Durbin, R., Eddy, S. R., Howe, K. L., Sonnhammer, E., 2000. The Pfam Protein Families Database. Nucl. Acids Res. 28 (1), 263–266.

Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., Wheeler, D. L., 2004. GenBank: update. Nucleic Acids Res 32 Database issue, 23–26.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., Bourne, P. E., 2000. The Protein Data Bank. Nucleic Acids Res 28 (1), 235–242.

Bottomley, M. J., Collard, M. W., Huggenvik, J. I., Liu, Z., Gibson, T. J., Sattler, M., 2001. The SAND domain structure defines a novel DNA-binding fold in transcriptional regulation. Nat Struct Biol 8 (7), 626–633.

Branden, C., Tooze, J. (Eds.), 1998. Introduction to Protein Structure, 2nd Edition. Garland Publishing Inc.

Bron, C., Kerbosch, J., 1973. Algorithm 457: finding all cliques of an undirected graph. Communications ACM 16, 575–577.

Cammer, S. A., Hoffman, B. T., Speir, J. A., Canady, M. A., Nelson, M. R., Knutson, S., Gallina, M., Baxter, S. M., Fetrow, J. S., 2003. Structure-based active site profiles for genome analysis and functional family subclassification. J Mol Biol 334 (3), 387–401.

Carraghan, R., Pardalos, P. M., 1990. An exact algorithm for the maximum clique problem. Operations Res, Lett. 9, 375–382.

Casari, G., Sander, C., Valencia, A., 1995. A method to predict functional residues in proteins. Nat Struct Biol 2 (2), 171–178.

Castagnetto, J. M., Hennessy, S. W., Roberts, V. A., Getzoff, E. D., Tainer, J. A., Pique, M. E., 2002. MDB: the Metalloprotein Database and Browser at The Scripps Research Institute. Nucleic Acids Res 30 (1), 379–382.

Chothia, C., 1992. Proteins. One thousand families for the molecular biologist. Nature 357 (6379), 543–544.

Cristobal, S., Zemla, A., Fischer, D., Rychlewski, L., Elofsson, A., 2001. A study of quality measures for protein threading models. BMC Bioinformatics 2 (1), 5, evaluation Studies.

Degtyarenko, K., Oct 2000. Bioinorganic motifs: towards functional classification of metalloproteins. Bioinformatics 16 (10), 851–864.

del Sol Mesa, A., Pazos, F., Valencia, A., 2003. Automatic methods for predicting functionally important residues. J Mol Biol 326 (4), 1289–1302.

DeLano, W., 2002. The PyMOL Molecular Graphics System.
URL `http://www.pymol.org`.

Enzyme Nomenclature, 1992. Recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the Nomenclature and Classification of Enzymes, NC-IUBMB.

Falquet, L., Pagni, M., Bucher, P., Hulo, N., Sigrist, C. J. A., Hofmann, K., Bairoch,

A., 2002. The PROSITE database, its status in 2002. Nucleic Acids Res 30 (1), 235–238.

Ferre, F., Ausiello, G., Zanzoni, A., Helmer-Citterich, M., 2004. SURFACE: a database of protein surface regions for functional annotation. Nucleic Acids Res 32 Database issue, 240–244.

Fetrow, J. S., Skolnick, J., 1998. Method for prediction of protein function from sequence using the sequence-to-structure-to-function paradigm with application to glutaredoxins/thioredoxins and T1 ribonucleases. J Mol Biol 281 (5), 949–968.

Fischer, D., Barret, C., Bryson, K., Elofsson, A., Godzik, A., Jones, D., Karplus, K. J., Kelley, L. A., MacCallum, R. M., Pawowski, K., Rost, B., Rychlewski, L., Sternberg, M., 1999. CAFASP-1: critical assessment of fully automated structure prediction methods. Proteins Suppl 3, 209–217.

Galkin, A., Sarikaya, E., Lehmann, C., Howard, A., Herzberg, O., 2003. Structure of YGFB from Haemophilus Influenzae (HI0817), a conserverd hypothtetical protein. to be published.

Galperin, M. Y., 2004. The Molecular Biology Database Collection: 2004 update. Nucleic Acids Res 32 Database issue, D3–22.

Gardiner, E. J., Artymiuk, P. J., Willett, P., 1997. Clique-detection algorithms

for matching three-dimensional molecular structures. J Mol Graph Model 15 (4), 245–253.

Gibson, T. J., Ramu, C., Gemund, C., Aasland, R., 1998. The APECED polyglandular autoimmune syndrome protein, AIRE-1, contains the SAND domain and is probably a transcription factor. Trends Biochem Sci 23 (7), 242–244.

Gregory, D. S., Martin, C. R., Cheetham, J. C., Rees, R. A., 1993. The prediction and characterization of metal binding sites in proteins. Pro. Eng. 6 (1), 29–35.

Hadley, C., Jones, D. T., 1999. A systematic comparison of protein structure classifications: SCOP, CATH and FSSP. Structure Fold Des 7 (9), 1099–1112.

Halford, S. E., Marko, J. F., 2004. How do site-specific DNA-binding proteins find their targets? Nucleic Acids Res 32 (10), 3040–3052.

Harris, M. A., Clark, J., Ireland, A., Lomax, J., Ashburner, M., Foulger, R., Eilbeck, K., Lewis, S., Marshall, B., Mungall, C., Richter, J., Rubin, G. M., Blake, J. A., Bult, C., Dolan, M., Drabkin, H., Eppig, J. T., Hill, D. P., Ni, L., Ringwald, M., Balakrishnan, R., Cherry, J. M., Christie, K. R., Costanzo, M. C., Dwight, S. S., Engel, S., Fisk, D. G., Hirschman, J. E., Hong, E. L., Nash, R. S., Sethuraman, A., Theesfeld, C. L., Botstein, D., Dolinski, K., Feierbach, B., Berardini, T., Mundodi, S., Rhee, S. Y., Apweiler, R., Barrell, D., Camon, E., Dimmer, E., Lee,

V., Chisholm, R., Gaudet, P., Kibbe, W., Kishore, R., Schwarz, E. M., Sternberg, P., Gwinn, M., Hannick, L., Wortman, J., Berriman, M., Wood, V., de la Cruz, N., Tonellato, P., Jaiswal, P., Seigfried, T., White, R., 2004. The Gene Ontology (GO) database and informatics resource. Nucleic Acids Res 32 Database issue, 258–261.

Harrison, S. C., 1991. A structural taxonomy of DNA-binding domains. Nature 353 (6346), 715–719.

Hendlich, M., Lackner, P., Weitckus, S., Floeckner, H., Froschauer, R., Gottsbacher, K., Casari, G., Sippl, M. J., 1990. Identification of native protein folds amongst a large number of incorrect models. The calculation of low energy conformations from potentials of mean force. J Mol Biol 216 (1), 167–180.

Henikoff, J. G., Greene, E. A., Pietrokovoski, S., Henikoff, S., 2000. Increased coverage of protein families with the Blocks Database servers. Nucl. Acids Res. 28 (1), 228–230.

Holm, L., Park, J., 2000. DaliLite workbench for protein structure comparison. Bioinformatics 16 (6), 566–567.

Holm, L., Sander, C., 1994. The FSSP database of structurally aligned protein fold families. Nucleic Acids Res 22 (17), 3600–3609.

Holm, L., Sander, C., 1998. Removing near-neighbour redundancy from large protein sequence collections. Bioinformatics 14 (5), 423–429.

Holm, R., Kennepohl, P., Solomon, E., 1996. Structural and Functional Aspects of Metal Sites in Biology. Chem Rev 96 (7), 2239–2314.

Humphrey, W., Dalke, A., Schulten, K., 1996. VMD - Visual Molecular Dynamic. J. Molec. Graphics 14 (1), 33–38.

Ihaka, R., Gentleman, R., 1996. R: A language for data analysis and graphics. Journal of Computational and Graphical Statistics 5 (3), 299–314.

Jones, D. T., 1999. GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. J Mol Biol 287 (4), 797–815.

Jones, D. T., 2001. Predicting novel protein folds by using FRAGFOLD. Proteins Suppl 5, 127–132.

Jones, D. T., McGuffin, L. J., 2003. Assembling novel protein folds from super-secondary structural fragments. Proteins 53 Suppl 6, 480–485.

Jones, D. T., Taylor, W. R., Thornton, J. M., 1992. A new approach to protein fold recognition. Nature 358 (6381), 86–89.

Jones, D. T., Ward, J. J., 2003. Prediction of disordered regions in proteins from position specific score matrices. Proteins 53 Suppl 6, 573–578.

Jones, S., Barker, J. A., Nobeli, I., Thornton, J. M., 2003a. Using structural motif templates to identify proteins with DNA binding function. Nucleic Acids Res 31 (11), 2811–2823.

Jones, S., Daley, D. T., Luscombe, N. M., Berman, H. M., Thornton, J. M., 2001. Protein-RNA interactions: a structural analysis. Nucleic Acids Res 29 (4), 943–954.

Jones, S., Shanahan, H. P., Berman, H. M., Thornton, J. M., 2003b. Using electrostatic potentials to predict DNA-binding sites on DNA-binding proteins. Nucleic Acids Res 31 (24), 7189–7198.

Jones, S., Thornton, J. M., 1997. Prediction of protein-protein interaction sites using patch analysis. J Mol Biol 272 (1), 133–143.

Jones, S., Thornton, J. M., 2004. Searching for functional sites in protein structures. Curr Opin Chem Biol 8 (1), 3–7.

Jones, S., van Heyningen, P., Berman, H. M., Thornton, J. M., 1999. Protein-DNA interactions: A structural analysis. J Mol Biol 287 (5), 877–896.

Kabsch, W., Sander, C., 1983. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22 (12), 2577–2637.

Karlin, S., Zhu, Z. Y., Karlin, K. D., 1997. The extended environment of mononuclear metal centers in protein structures. Proc. Natl. Acad. Sci. 94 (26), 14225–14230.

Kelley, L. A., MacCallum, R. M., Sternberg, M. J., 2000. Enhanced genome annotation using structural profiles in the program 3D-PSSM. J Mol Biol 299 (2), 499–520.

Kinoshita, K., Nakamura, H., 2003. Identification of protein biochemical functions by similarity search using the molecular surface database eF-site. Protein Sci 12 (8), 1589–1595.

Kleywegt, G. J., 1999. Recognition of spatial motifs in protein structures. J Mol Biol 285 (4), 1887–1897.

Kopp, J., Schwede, T., 2004. The SWISS-MODEL Repository of annotated three-dimensional protein structure homology models. Nucleic Acids Res 32 Database issue, 230–234.

Laskowski, R. A., 1995. SURFNET: a program for visualizing molecular surfaces, cavities, and intermolecular interactions. J Mol Graph 13 (5), 323–330.

Laskowski, R. A., 2001. PDBsum: summaries and analyses of PDB structures. Nucleic Acids Res 29 (1), 221–222.

Laskowski, R. A., Hutchinson, E. G., Michie, A. D., Wallace, A. C., Jones, M. L., Thornton, J. M., 1997. PDBsum: a Web-based database of summaries and analyses of all PDB structures. Trends Biochem Sci 22 (12), 488–490.

Laskowski, R. A., Watson, J. D., Thornton, J. M., 2003. From protein structure to biochemical function? J Struct Funct Genomics 4 (2-3), 167–177.

Lee, B., Richards, F. M., 1971. The interpretation of protein structures: estimation of static accessibility. J Mol Biol 55 (3), 379–400.

Liang, M. P., Banatao, D. R., Klein, T. E., Brutlag, D. L., Altman, R. B., 2003a. WebFEATURE: An interactive web tool for identifying and visualizing functional sites on macromolecular structures. Nucleic Acids Res 31 (13), 3324–3327.

Liang, M. P., Brutlag, D. L., Altman, R. B., 2003b. Automated construction of structural motifs for predicting functional sites on protein structures. Pac Symp Biocomput, 204–215.

Lichtarge, O., Yao, H., Kristensen, D. M., Madabushi, S., Mihalek, I., 2003. Accurate and scalable identification of functional sites by evolutionary tracing. J Struct Funct Genomics 4 (2-3), 159–166.

Lo Conte, L., Ailey, B., Hubbard, T. J., Brenner, S. E., Murzin, A. G., Chothia, C.,

2000. SCOP: a structural classification of proteins database. Nucleic Acids Res 28 (1), 257–259.

Luscombe, N. M., Austin, S. E., Berman, H. M., Thornton, J. M., 2000. An overview of the structures of protein-DNA complexes. Genome Biol 1 (1), REVIEWS001.

Luscombe, N. M., Laskowski, R. A., Thornton, J. M., 1997. NUCPLOT: a program to generate schematic diagrams of protein-nucleic acid interactions. Nucleic Acids Res 25 (24), 4940–4945.

Luscombe, N. M., Thornton, J. M., 2002. Protein-DNA interactions: amino acid conservation and the effects of mutations on binding specificity. J Mol Biol 320 (5), 991–1009.

McCulloch, W. S., Pitts, W., 1943. "A logical calculus of ideas immanent in nervous activity". Bull. Math. Biophys. 5, 115–133.

McGuffin, L. J., Bryson, K., Jones, D. T., 2000. The PSIPRED protein structure prediction server. Bioinformatics 16 (4), 404–405.

McGuffin, L. J., Bryson, K., Jones, D. T., 2001. What are the baselines for protein fold recognition? Bioinformatics 17 (1), 63–72.

McGuffin, L. J., Jones, D. T., 2003. Improvement of the GenTHREADER method for genomic fold recognition. Bioinformatics 19 (7), 874–881.

McGuffin, L. J., Street, S., Sorensen, S.-A., Jones, D. T., 2004a. The genomic threading database. Bioinformatics 20 (1), 131–132.

McGuffin, L. J., Street, S. A., Bryson, K., Sorensen, S.-A., Jones, D. T., 2004b. The Genomic Threading Database: a comprehensive resource for structural annotations of the genomes from key organisms. Nucleic Acids Res 32 Database issue, 196–199.

McLaughlin, W. A., Berman, H. M., 2003. Statistical models for discerning protein structures containing the DNA-binding helix-turn-helix motif. J Mol Biol 330 (1), 43–55.

Moult, J., Fidelis, K., Zemla, A., Hubbard, T., 2001. Critical assessment of methods of protein structure prediction (CASP): round IV. Proteins Suppl 5, 2–7.

Moult, J., Hubbard, T., Bryant, S. H., Fidelis, K., Pedersen, J. T., 1997. Critical assessment of methods of protein structure prediction (CASP): round II. Proteins Suppl 1, 2–6.

Moult, J., Melamud, E., 2000. From fold to function. Curr Opin Struct Biol 10 (3), 384–389.

Mulder, N. J., Apweiler, R., Attwood, T. K., Bairoch, A., Bateman, A., Binns, D., Biswas, M., Bradley, P., Bork, P., Bucher, P., Copley, R., Courcelle, E.,

Durbin, R., Falquet, L., Fleischmann, W., Gouzy, J., Griffith-Jones, S., Haft, D., Hermjakob, H., Hulo, N., Kahn, D., Kanapin, A., Krestyaninova, M., Lopez, R., Letunic, I., Orchard, S., Pagni, M., Peyruc, D., Ponting, C. P., Servant, F., Sigrist, C. J. A., 2002. InterPro: an integrated documentation resource for protein families, domains and functional sites. Brief Bioinform 3 (3), 225–235.

Nadassy, K., Tomas-Oliveira, I., Alberts, I., Janin, J., Wodak, S. J., 2001. Standard atomic volumes in double-stranded DNA and packing in protein–DNA interfaces. Nucleic Acids Res 29 (16), 3362–3376.

Nadassy, K., Wodak, S. J., Janin, J., 1999. Structural features of protein-nucleic acid recognition sites. Biochemistry 38 (7), 1999–2017.

Nagano, N., Hutchinson, E. G., Thornton, J. M., 1999. Barrel structures in proteins: automatic identification and classification including a sequence analysis of TIM barrels. Protein Sci 8 (10), 2072–2084.

Nature Editorial, 2004. PSI-phase 1 and beyond.

Needleman, S. B., Wunsch, C. D., 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol 48 (3), 443–453.

Orengo, C. A., Bray, J. E., Buchan, D. W. A., Harrison, A., Lee, D., Pearl, F.

M. G., Sillitoe, I., Todd, A. E., Thornton, J. M., 2002. The CATH protein family database: a resource for structural and functional annotation of genomes. Proteomics 2 (1), 11–21.

Orengo, C. A., Jones, D. T., Thornton, J. M., 1994. Protein superfamilies and domain superfolds. Nature 372 (6507), 631–634.

Orengo, C. A., Jones, D. T., Thornton, J. M. (Eds.), 2003. Bioinformatics: Genes, Protiens and Computers. BIOS Scientific Publishers Ltd.

Orengo, C. A., Taylor, W. R., 1996. SSAP: sequential structure alignment program for protein structure comparison. Methods Enzymol 266, 617–635.

Orengo, C. A., Todd, A. E., Thornton, J. M., 1999. From protein structure to function. Curr Opin Struct Biol 9 (3), 374–382.

Porter, C. T., Bartlett, G. J., Thornton, J. M., 2004. The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data. Nucleic Acids Res 32 Database issue, 129–133.

Przytycka, T., Aurora, R., Rose, G. D., 1999. A protein taxonomy based on secondary structure. Nat Struct Biol 6 (7), 672–682.

Riedmiller, M., Braun, H., 1993. A direct adaptive method for faster backpropagation learning: The RPROP algorithm. In: Proc. of the IEEE Intl. Conf. on Neural

Networks. San Francisco, CA, pp. 586–591.

URL `citeseer.ist.psu.edu/riedmiller93direct.html`

Rost, B., Sander, C., 1994. Conservation and prediction of solvent accessibility in protein families. Proteins 20 (3), 216–226.

Russell, R. B., Sasieni, P. D., Sternberg, M. J., 1998. Supersites within superfolds. Binding site similarity in the absence of homology. J Mol Biol 282 (4), 903–918.

Rychlewski, L., Fischer, D., Elofsson, A., 2003. LiveBench-6: large-scale automated evaluation of protein structure prediction servers. Proteins 53 Suppl 6, 542–547.

Sali, A., Potterton, L., Yuan, F., van Vlijmen, H., Karplus, M., 1995. Evaluation of comparative protein modeling by MODELLER. Proteins 23 (3), 318–326.

Schmitt, S., Kuhn, D., Klebe, G., 2002. A new method to detect related function among proteins independent of sequence and fold homology. J Mol Biol 323 (2), 387–406.

Shanahan, H. P., Garcia, M. A., Jones, S., Thornton, J. M., 2004. Identifying DNA-binding proteins using structural motifs and the electrostatic potential. Nucleic Acids Res 32 (16), 4732–4741.

Siew, N., Elofsson, A., Rychlewski, L., Fischer, D., 2000. MaxSub: an automated

measure for the assessment of protein structure prediction quality. Bioinformatics 16 (9), 776–785.

Simons, K. T., Bonneau, R., Ruczinski, I., Baker, D., 1999. Ab initio protein structure prediction of CASP III targets using ROSETTA. Proteins Suppl 3, 171–176.

Sodhi, J. S., Bryson, K., McGuffin, L. J., Ward, J. J., Wernisch, L., Jones, D. T., 2004. Predicting metal-binding site residues in low-resolution structural models. J Mol Biol 342 (1), 307–320.

Stark, A., Sunyaev, S., Russell, R. B., 2003. A model for statistical significance of local similarities in structure. J Mol Biol 326 (5), 1307–1316.

Stawiski, E. W., Gregoret, L. M., Mandel-Gutfreund, Y., 2003. Annotating nucleic acid-binding function based on protein structure. J Mol Biol 326 (4), 1065–1079.

Tatusov, R. L., Koonin, E. V., Lipman, D. J., 1997. A genomic perspective on protein families. Science 278 (5338), 631–637.

Teng, T. Y., 1990. Mounting of crystals for macromolecular crystallography in a free-standing thin film. J. Appl. Cryst. 12, 387–391.

Thore, S., Mauxion, F., Seraphin, B., Suck, D., 2003. X-ray structure and activity of the yeast Pop2 protein: a nuclease subunit of the mRNA deadenylase complex. EMBO Rep. 12 (4), 1150–1155.

Wallace, A. C., Borkakoti, N., Thornton, J. M., 1997. TESS: a geometric hashing algorithm for deriving 3D coordinate templates for searching structural databases. Application to enzyme active sites. Protein Sci 6 (11), 2308–2323.

Wallace, A. C., Laskowski, R. A., Thornton, J. M., 1995. LIGPLOT: a program to generate schematic diagrams of protein-ligand interactions. Protein Eng 8 (2), 127–134.

Wallner, B., Elofsson, A., 2003. Can correct protein models be identified? Protein Sci 12 (5), 1073–1086.

Wang, J., Sykes, B. D., Ryan, R., 2002. Structural basis for the conformational adaptability of apolipophorin III, a helix-bundle exchangeable apolipoprotein. Proc Natl Acad Sci 99 (3), 1188–1193.

Ward, J. J., Sodhi, J. S., McGuffin, L. J., Buxton, B. F., Jones, D. T., 2004. Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. J Mol Biol 337 (3), 635–645.

Wei, L., Altman, R. B., 1998. Recognizing protein binding sites using statistical descriptions of their 3D environments. Pac Symp Biocomput, 497–508.

Wei, L., Altman, R. B., 2003. Recognizing complex, asymmetric functional sites in

protein structures using a Bayesian scoring function. J Bioinform Comput Biol 1 (1), 119–138.

Wei, L., Huang, E. S., Altman, R. B., 1999. Are predicted structures good enough to preserve functional sites? Structure Fold Des 7 (6), 643–650.

Westbrook, J., Feng, Z., Jain, S., Bhat, T. N., Thanki, N., Ravichandran, V., Gilliland, G. L., Bluhm, W., Weissig, H., Greer, D. S., Bourne, P. E., Berman, H. M., 2002. The Protein Data Bank: unifying the archive. Nucleic Acids Res 30 (1), 245–248.

Yao, H., Kristensen, D. M., Mihalek, I., Sowa, M. E., Shaw, C., Kimmel, M., Kavraki, L., Lichtarge, O., 2003. An accurate, sensitive, and scalable method to identify functional sites in protein structures. J Mol Biol 326 (1), 255–261.

Zhang, L., Godzik, A., Skolnick, J., Fetrow, J. S., 1998. Functional analysis of the Escherichia coli genome for members of the alpha/beta hydrolase family. Fold Des 3 (6), 535–548.

Appendix A

Additional FuncSite Analysis

Amino Acid Distribution

Zinc

Figure A.1 illustrates the amino acid distribution for the zinc containing sites. Cysteine residues are by far the most prevalent within 5Å comprising 40.5% as compared to only 3% for the non-site data derived from the zinc dataset. Histidine is present in 14% of residues as compared to 2.5% of non-site data whilst aspartate residues are also commonly found with 16% as compared to only 1.6% in the non-site data. Analysis of the medium range residues reveals a striking abundance of His (22%) highlighting the layered nature known to exist in the environment surrounding metal ion sites needed for selective binding. These observations are consistent with the

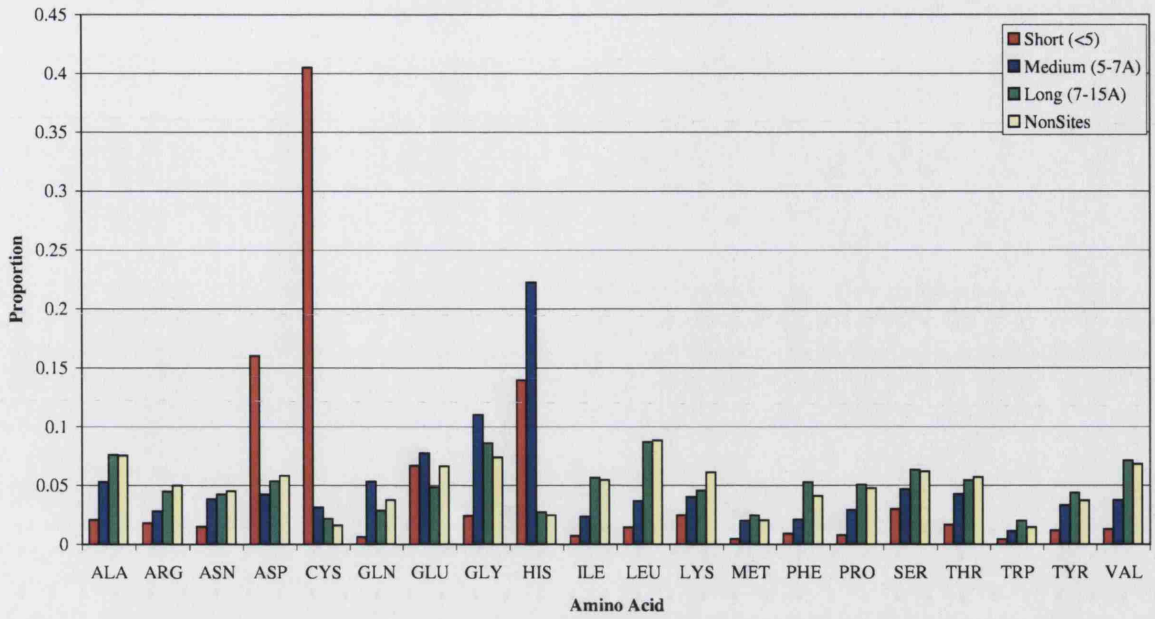


Figure A.1: Zinc Sites Amino Acid Distribution

fact that zinc atoms preferred affinity to sulphur atoms of Cys, imidazole nitrogen of His as well as carboxylate oxygen atoms of aspartate.

Magnesium

Aspartate is also observed as the preferred residue type around Mg^{2+} ions comprising 30.4% as compared to only 5% for non-sites. Intrestingly both Glu and Gly are seen more frequently in the medium range (5-7Å) interactions, 14% and 11% respectively (Figure A.2). Another intriguing observation is the peak of Ser and Thr, this feature is unique to the Mg^{2+} site set occurring in 10% and 8% of residues in Mg^{2+} sites

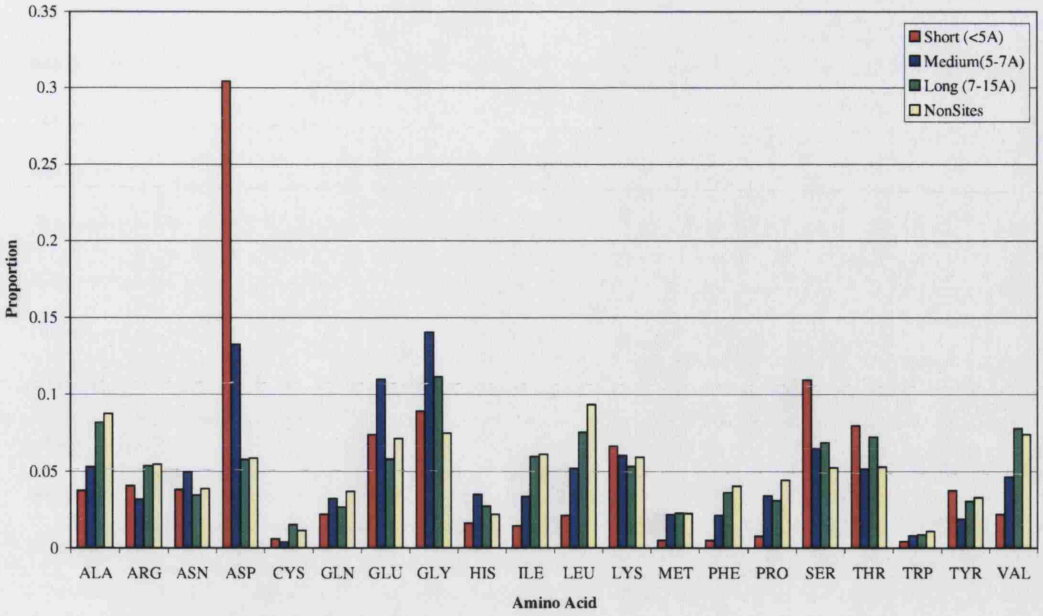


Figure A.2: Magnesium Sites Amino Acid Distribution

and indicating side chain oxygen atoms from these residues also contribute to ion ligation.

Copper

The copper sites data (Figure A.3) reveals a similar pattern to zinc sites data in that both His and Cys are more prevalent around the ion site. However, unlike zinc the copper data suggests that imidazole nitrogen from His residues are more important in ligation comprising 46% (3% for non-site). Cys residues are also more common around Cu^{2+} consiting 17% of site residues (1% for non-site). Intrestingly the role of Met sulphur atoms are also highlighted in Cu^{2+} sites where 11.3% are found as

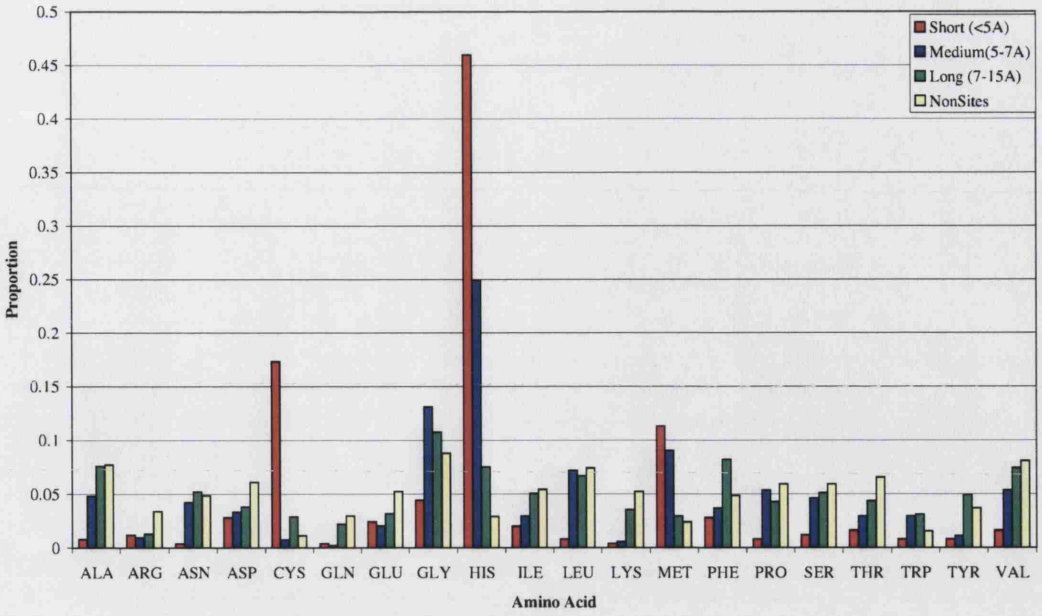


Figure A.3: Copper Sites Amino Acid Distribution

oppose to only 2% in non-sites.

Iron

The most abundant residues observed within 5Å of iron ions were Cys (18%) and Asp (12.5%). However, His residues appear more frequently within the medium range cut-off consisting of 30.4% as compared to 5% and 2% for residues $\leq$5Å and non-site regions (Figure A.4). This again illustrates the longer-range interactions that His imidazole nitrogen participate in and is likely to be an important contributor to the secondary shell environment allowing selective metal binding.

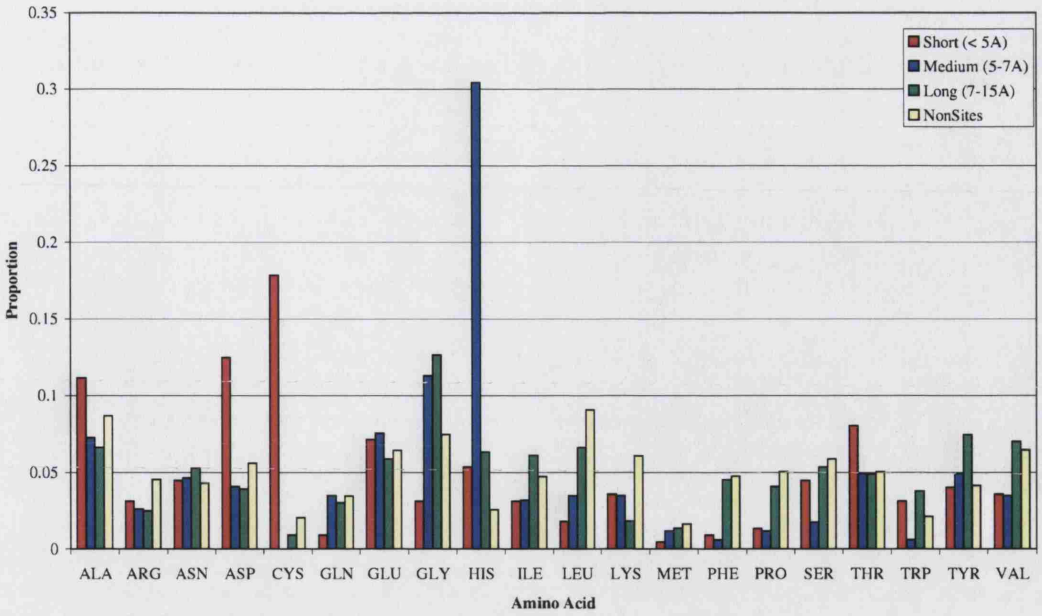


Figure A.4: Iron Sites Amino Acid Distribution

Secondary Structure Analysis

Metal/SS	Distance from Metal			
Calcium	**≤5Å**	**5-7 Å**	**7-15 Å**	**>15Å(Non-Sites)**
hline *Sheet*	24.8%	24.8%	32.1%	24.4%
Helix	23.5%	27.2%	39%	30.4%
Coil	51.7%	48%	29%	45.2%
Zinc	**≤5Å**	**5-7 Å**	**7-15 Å**	**>15Å(Non-Sites)**
Sheet	21%	25%	31.7%	21.4%
Helix	22.1%	37.9%	36.8%	34.09%
Coil	56.9%	37.1%	31.4%	44.5%
Magnesium	**≤5Å**	**5-7 Å**	**7-15 Å**	**>15Å(Non-Sites)**
Sheet	20.4%	20.8%	24.8%	19.39%
Helix	33.4%	24.2%	32%	40.19%
Coil	46.2%	54.9%	43.15%	40.41%
Copper	**≤5Å**	**5-7 Å**	**7-15 Å**	**>15Å(Non-Sites)**
Sheet	26.6%	33.6%	34.9%	30.8%
Helix	8.9%	14%	19.4%	22%
Coil	64.5%	52.4%	45.7%	47.2%
Manganese	**≤5Å**	**5-7 Å**	**7-15 Å**	**>15Å(Non-Sites)**
Sheet	24.4%	27.6%	30%	20.8%
Helix	18.4%	20.22%	25.45%	37.5%
Coil	53.17%	52.15%	44.4%	41.7%
Iron	**≤5Å**	**5-7 Å**	**7-15 Å**	**>15Å(Non-Sites)**
Sheet	12.9%	19.13%	23.04%	15.8%
Helix	29.01%	47.53%	40.43%	41.6%
Coil	58.03%	33.3%	36.52%	42.5%

Table A.1: Metal Sites Secondary Structure Analysis

Appendix B

Neural Networks and Backpropagation Training

The growing biological databases, sequence and structural, provide a rich source of information from which rules and patterns can be identified. Briefly, learning algorithms aim to improve the performance of a task on the basis of previous experience. Such approaches have been successfully applied to variety of problem domains and are particularly powerful for the type of pattern rich data within Bioinformatics. An in-depth review of machine learning is beyond the scope of the current discussion, however, a brief overview of a commonly used method is presented.

Artificial Neural Networks

Artificial neural networks (ANN) are particularly popular in Bioinformatics. The technique has been applied to a wide variety of different problems ranging from secondary structure prediction through to the analysis of gene expression data. The design of neural networks was originally inspired by the mechanisms by which neurons function in the human brain (McCulloch and Pitts, 1943). Essentially, knowledge is acquired by a learning process which is achieved through adjusting weights which connect virtual neurons.

The topology of a computational neural network is linked to the learning algorithm used in training. A particularly popular, and effective example, is the multilayer feed-forward acrhitecture. Such ANN employ a hidden layer located between the input and output layers of the network. The added dimensionality allows higher

order relationships between input values to be managed more effectively. This is particularly important when the number of inputs to the network is large.

Supervised learning algorithms modify the weights of a ANN in order to reach a desired objective. This leads to the most important features of any machine learning method: the ability to generalise, beyond the set of examples used in training, to new data. A common cause for poor generalisation is observed when the network essentially memorises the examples used in training (over-fitting) and is unable to effectively map new input data to output space. Factors which influence this behaviour include the complexity of the problem being investigated, network topology, and the number of examples in training.

Cross-validation is the means by which examples are re-sampled during training. This often involves partitioning the data into training, testing and in some cases validations sets. Training is performed using the training partition and the network is evaluated on the testing set. An important aspect, particularly for Bioinformatics data, is to ensure no trivial relationships (such as two proteins with high sequence identity) are presented in testing and training. Such cases will result in overly optimistic predictions and mask the true generalisation qualities of the network.

A commonly used strategy is k-fold cross-validation where the training example are partitioned into a number (k) of sets. Training is performed leaving out a single set in turn which is used for testing. At the end of training the results from all the

testing sets are pooled to provide an overall assessment. In early stopping a sub-set of the training examples are used as a validation set: the error is calculated for this set and training is ceased when the error reaches a defined threshold.

Increasing the number of inputs increases the number of dimensions. This is known as the curse of dimensionality. Although multi-layer peceptrons (MLP) are less sensitive to this problem, as compared to other network topologies, it can still lead to a loss of generalisation abilities. Fortunately there are several techniques which can be used to prevent such deleterious effects. One such approach, which can be easily implemented, is to simply reduce the number of weights by removing inputs or nodes in the hidden layer. Another approach is to train the neural network using early stopping whereby training is ceased upon reaching a pre-defined error theshold on an independent validation data set.

Backpropagation Training

The backpropagation learning algorithm is the most commonly used supervised training methods for neural networks. Briefly, the objective is to locate a suitable error minimum by iteratively reducing the error of predictions by comparing current network outputs to the desired goal.

Many variations of backpropagation exist. The Matlab neural network toolbox (The MathWorks Inc., MA, USA) provides access to a number of varying approaches.

One of the simplest approaches is that of gradient decent whereby the network weights are updated in the direction of the minimum error.

The mean error of a neural network (E) gives a measure of the network performance over the entire training dataset (N) and is defined as the average squared error between the output of the network and the target value as follows:

$$\Delta E = \frac{1}{N}\sum_{i=1}^{n}(t-o)^2 \tag{B.1}$$

The backpropagation training algorithm adjusts the weights of the network such that E moves towards the minimum negative gradient. Thus the algorithm can be defined as a gradient decent approach to locate the global minimum of the error surface.

Thus the error is calculated for network with initialised weights, this error is reduced by calculating a weight update term which is proportional to the error gradient:

$$\frac{\delta E}{\delta w_{ij}} = \frac{\delta E}{\delta o_j}\frac{\delta o_j}{\delta a_j}\frac{\delta a_j}{\delta w_{ij}} \tag{B.2}$$

where w_{ji} is the weight connecting neuron i to j, o_j is the actual output of neuron j and a_j is the activation of j given by the sum of weights x inputs feeding j. Using the chain rule the local error gradient is given by:

$$\frac{\delta E}{\delta w_{ij}} = -(d_j - o_j) f'(a_j) x_i \tag{B.3}$$

where $f'(a_j)$ = derivative of actual output of neuron j with respect to activation of j, d_j is the desired output of j and x_i is the input to j. Thus the correction function is given by delta rule:

$$\Delta w_{ij} = -\eta \frac{\delta E}{\delta w_{ij}} \tag{B.4}$$

where η = is the learning rate that is the step width of each iteration on the error surface and δj is $-(d_j - o_j) f'(a_j)$. Thus the new weight is given by the sum of the delta rule and existing weight.

Appendix C

Publications Arising from this Thesis

1. Sodhi, J.S., Bryson, K., McGuffin, L.J., Ward, J.J., Wernisch, L., Jones, D.T. 2004. Prediction Metal-binding Site Residues in Low-Resolution Structural Models. J. Mol. Biol. 342 (1), 307-320.

2. Sodhi, J.S., Bryson, K., McGuffin, L.J., Ward, J.J., Wernisch, L., Jones, D.T. 2004. Automatic Prediction of Functional Site Regions in Low-Resolution Protein Structure. IEEE proceedings, CSB 2004: 702-703.